DOCUMENT RESUME

ED 328 559 TM 014 727

AUTHOR Heyneman, Stephen P., Ed.; Fagerlind, Ingemar, Ed.
TITLE University Examinations and Standardized Testing: Principles, Experience, and Policy Options. World Bank Technical Paper Number 78. Proceedings of a Seminar on the Uses of Standardized Tests and Selection Examinations (Beijing, China, April 1985).
INSTITUTION World Bank, Washington, D. C.
REPORT NO ISBN-0-8213-0990-0
PUB DATE Jan 88
NOTE 235p.
AVAILABLE FROM Publications Sales Unit, Department F, The World Bank, 1818 H Street, N.W., Washington, DC 20433.
PUB TYPE Collected Works - Conference Proceedings (021)

EDRS PRICE MF01 Plus Postage. PC Not Available from EDRS.
DESCRIPTORS Academic Achievement; Admission Criteria; College Admission; *College Entrance Examinations; Comparative Analysis; Cross Cultural Studies; Developing Nations; Educational Assessment; Educational Improvement; *Educational Policy; Educational Quality; English; Foreign Countries; Higher Education; *Standardized Tests; Test Construction; *Testing Programs; Test Use; *Universities
IDENTIFIERS Australia; *China; England; Japan; National Assessment of Educational Progress; Sweden; United States

ABSTRACT

In September 1984, the Chinese government asked the Economic Development Institute of the World Bank to assist the officials of the Chinese Ministry of Education in thinking through some policy options for examinations and standardized testing. This document summarizes the descriptions of testing programs and advice provided to these Chinese officials at a meeting held in April 1985. In addition to an introduction by S. P. Heyneman and I. Fagerlind, the following papers are provided: (1) "Admission to Higher Education in Japan" (T. Hidano); (2) "Examinations for University Selection in England" (J. L. Reddaway); (3) "Admission to Higher Education in the United States: The Role of the Educational Testing Service" (R. J. Solomon); (4) "Public Examinations in Australia" (J. P. Keeves); (5) "Education in Sweden: Assessment of Student Achievement and Selection for Higher Education" (S. Marklund); (6) "A Brief Introduction to the System of Higher School Enrollment Examinations in China" (L. Zhen); (7) "Designing the English Language Proficiency Test in China" (G. Shichun); (8) "Assessing the Quality of Education over Time: The Role of the National Assessment of Educational Progress (NAEP)" (A. E. LaPointe); (9) "Cross-National Comparisons in Educational Achievement: The Role of the International Association for the Evaluation of Educational Achievement (IEA)" (J. P. Keeves); (10) "Examinations as an Instrument To Improve Pedagogy" (A. Somerset); and (11) "Improving University Selection, Educational Research, and Educational Management in Developing Countries: The Role of Examinations and Standardized Testing" (S. P. Heyneman). Collectively, the papers contain 31 tables and 13 figures. (SLD)

ED328559

SCOPE OF INTEREST NOTICE

The ERIC Facility has assigned this document for processing to: TM

In our judgment, this document is also of interest to the Clearinghouses noted to the right. Indexing should reflect their special points of view. HE

WORLD BANK TECHNICAL PAPER NUMBER 78

University Examinations and Standardized Testing

Principles, Experience, and Policy Options

Stephen P. Heyneman and
Ingemar Fägerlind, editors

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

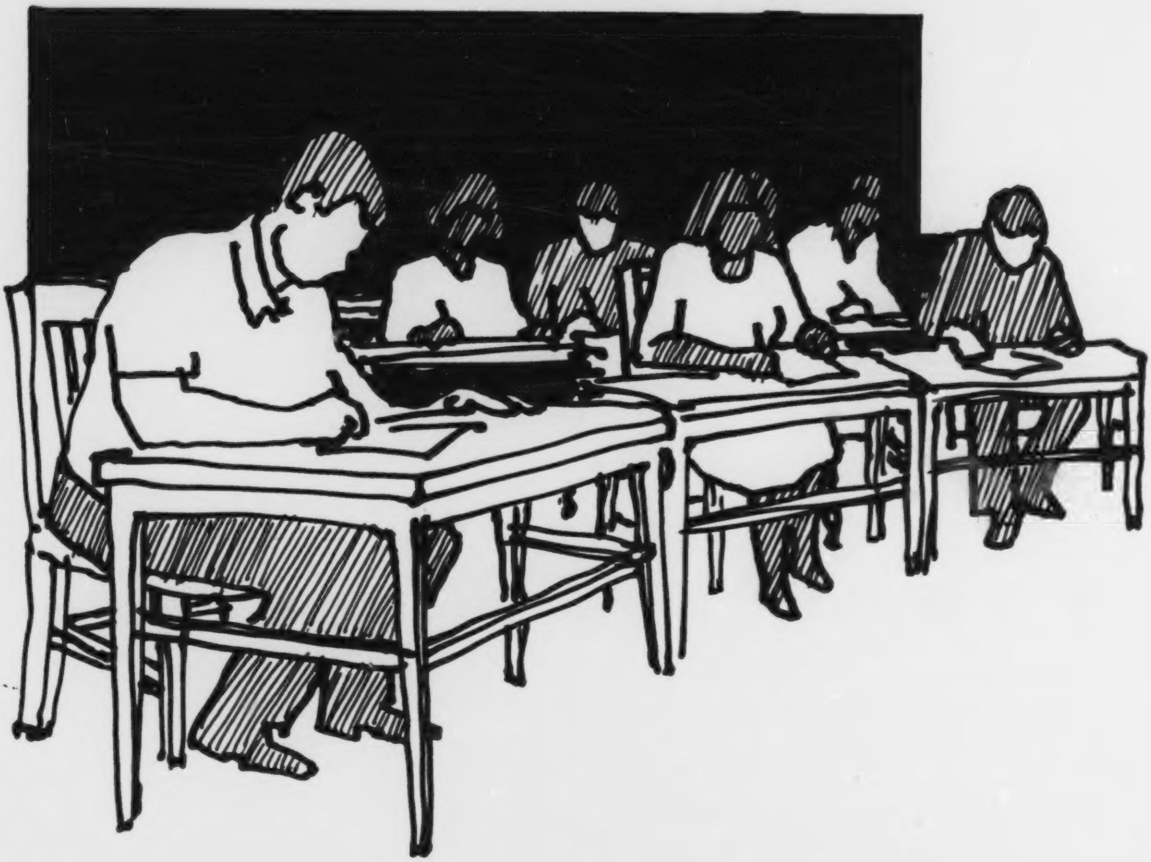
☐ Minor changes have been made to improve reproduction quality.

- Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL IN MICROFICHE ONLY HAS BEEN GRANTED BY

J. FEATHER

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."



TM014727

RECENT WORLD BANK TECHNICAL PAPERS

No. 20. Water Quality in Hydroelectric Projects: Considerations for Planning in Tropical Forest Regions

No. 21. Industrial Restructuring: Issues and Experiences in Selected Developed Economies

No. 22. Energy Efficiency in the Steel Industry with Emphasis on Developing Countries

No. 23. The Twinning of Institutions: Its Use as a Technical Assistance Delivery System

No. 24. World Sulphur Survey

No. 25. Industrialization in Sub-Saharan Africa: Strategies and Performance
(also in French, 25F)

No. 26. Small Enterprise Development: Economic Issues from African Experience
(also in French, 26F)

No. 27. Farming Systems in Africa: The Great Lakes Highlands of Zaire, Rwanda, and Burundi
(also in French, 27F)

No. 28. Technical Assistance and Aid Agency Staff: Alternative Techniques for Greater Effectiveness

No. 29. Handpumps Testing and Development: Progress Report on Field and Laboratory Testing

No. 30. Recycling from Municipal Refuse: A State-of-the-Art Review and Annotated Bibliography

No. 31. Remanufacturing: The Experience of the United States and Implications for Developing Countries

No. 32. World Refinery Industry: Need for Restructuring

No. 33. Guidelines for Calculating Financial and Economic Rates of Return for DFC Projects
(also in French, 33F, and Spanish, 33S)

No. 34. Energy Efficiency in the Pulp and Paper Industry with Emphasis on Developing Countries

No. 35. Potential for Energy Efficiency in the Fertilizer Industry

No. 36. Aquaculture: A Component of Low Cost Sanitation Technology

No. 37. Municipal Waste Processing in Europe: A Status Report on Selected Materials and Energy Recovery Projects

No. 38. Bulk Shipping and Terminal Logistics

No. 39. Cocoa Production: Present Constraints and Priorities for Research

No. 40. Irrigation Design and Management: Experience in Thailand

No. 41. Fuel Peat in Developing Countries

No. 42. Administrative and Operational Procedures for Programs for Sites and Services and Area Upgrading

No. 43. Farming Systems Research: A Review

No. 44. Animal Health Services in Sub-Saharan Africa: Alternative Approaches

No. 45. The International Road Roughness Experiment: Establishing Correlation and and a Calibration Standard for Measurements

No. 46. Guidelines for Conducting and Calibrating Road Roughness Measurements

No. 47. Guidelines for Evaluating the Management Information Systems of Industrial Enterprises

No. 48. Handpumps Testing and Development: Proceedings of a Workshop in China

No. 49. Anaerobic Digestion: Principals and Practices for Biogas Systems

No. 50. Investment and Finance in Agricultural Service Cooperatives

(List continues on the inside back cover.)

WORLD BANK TECHNICAL PAPER NUMBER 78

University Examinations and Standardized Testing

Principles, Experience, and Policy Options

Stephen P. Heyneman and Ingemar Fägerlind, editors

The World Bank
Washington, D.C.

Copyright © 1988
The International Bank for Reconstruction
and Development/THE WORLD BANK
1818 H Street, N.W.
Washington, D.C. 20433, U.S.A.

All rights reserved
Manufactured in the United States of America
First printing January 1988

Technical Papers are not formal publications of the World Bank, and are circulated to encourage discussion and comment and to communicate the results of the Bank's work quickly to the development community; citation and the use of these papers should take account of their provisional character. The findings, interpretations, and conclusions expressed in this paper are entirely those of the author(s) and should not be attributed in any manner to the World Bank, to its affiliated organizations, or to members of its Board of Executive Directors or the countries they represent. Any maps that accompany the text have been prepared solely for the convenience of readers; the designations and presentation of material in them do not imply the expression of any opinion whatsoever on the part of the World Bank, its affiliates, or its Board or member countries concerning the legal status of any country, territory, city, or area or of the authorities thereof or concerning the delimitation of its boundaries or its national affiliation.

Because of the informality and to present the results of research with the least possible delay, the typescript has not been prepared in accordance with the procedures appropriate to formal printed texts, and the World Bank accepts no responsibility for errors.

The most recent World Bank publications are described in the catalog *New Publications*, a new edition of which is issued in the spring and fall of each year. The complete backlist of publications is shown in the annual *Index of Publications*, which contains an alphabetical title list and indexes of subjects, authors, and countries and regions; it is of value principally to libraries and institutional purchasers. The latest edition of each of these is available free of charge from the Publications Sales Unit, Department F, The World Bank, 1818 H Street, N.W., Washington, D.C. 20433, U.S.A., or from Publications, The World Bank, 66, avenue d'Iéna, 75116 Paris, France.

Stephen P. Heyneman is chief of the Education and Training Design Division of the Economic Development Institute at the World Bank. Ingemar Fägerlind is director of the Institute of International Education at the University of Stockholm and a consultant to the World Bank.

Library of Congress Cataloging-in-Publication Data

University examinations and standardized testing.

(World Bank technical paper, ISSN 0253-7494 ; no. 78)
Bibliography: p.
1. Universities and colleges--Entrance requirements.
2. Universities and colleges--Entrance examinations.
I. Heyneman, Stephen P. II. Fägerlind, Ingemar,
1935- . III. Series.
LB2351.U59 1988 378'.1057 88-38
ISBN 0-8213-0990-0

ABSTRACT

As countries today renew their use of educational testing, new concerns have arisen about how better to manage such testing. China provides one example of the new emphasis on purposefully managing the policies toward educational testing. In September 1984, the Chinese government asked the Economic Development Institute of the World Bank to assist the officials of the Ministry of Education in thinking through some of the policy options for examinations and standardized testing. As a result, a meeting was held in April 1985. This book summarizes the descriptions of testing systems in selected OECD countries and the advice given to the government officials following the meeting. The attention devoted to problems of logistics and to economies of scale are perhaps more pertinent to large, heterogeneous countries such as China. But this book contends that many of the principles discussed at the meeting and presented here are applicable to developing countries generally.

CONTRIBUTORS

Ingemar Fägerlind, Director, Institute of International Education, University of Stockholm, Sweden

Gui Shichun, Professor of English, Guangzhou Institute of Foreign Languages, Guangzhou, China

Stephen P. Heyneman, Chief, Education and Training Design Division, Economic Development Institute, The World Bank

Tadashi Hidano, Deputy Director-General, National Center for University Entrance Examination, Tokyo, Japan

John Philip Keeves, Centre for the Study of Higher Education, University of Melbourne, Parkville, Victoria, Australia

Archie E. LaPointe, Executive Director, National Assessment of Educational Progress, Princeton, New Jersey, United States

Lu Zhen, Deputy Director, China International Examinations Coordination Bureau, Ministry of Education, Beijing, China

Sixten Marklund, Professor, Institute of International Education, University of Stockholm, Sweden

John Lewis Reddaway, Secretary, University of Cambridge Local Examinations Syndicate, Cambridge, United Kingdom

Robert Solomon, Executive Vice President, Educational Testing Service, Princeton, New Jersey, United States

Anthony Somerset, Institute of Development Studies, University of Sussex, Sussex, United Kingdom

CONTENTS

ABSTRACT iii

PREFACE xi

I. INTRODUCTION Stephen P. Heyneman and Ingemar Fägerlind 1

II. TESTING FOR SELECTION TO UNIVERSITY 7

1. ADMISSION TO HIGHER EDUCATION IN JAPAN Tadashi Hidano 9

The History of Examinations for Admission to Higher Education 9
The National Center for University Entrance Examination (NCUEE) 12
Setting and Reviewing the Joint First-Stage Achievement Test (JFSAT) 17
University Selection Procedure 20
Research 22
Technical and Social Issues 24

2. EXAMINATIONS FOR UNIVERSITY SELECTION IN ENGLAND John Lewis Reddaway 26

Historical Origin and Functions 26
Organization of the General Certificate of Education (GCE) 27
The University of Cambridge Local Examinations Syndicate (UCLES) 28
UCLES GCE Advanced-Level Technology 29
Marking the Examination 31
Contrast with Other Systems 32
GCE and University Selection 33
Appendix 1. The Syndicate and Its Committees 35
Appendix 2. Technology, GCE Advanced Level 36
Appendix 3. Technology, GCE Advanced Level, Paper 1 43
Appendix 4. Technology, GCE Advanced Level, Paper 2 49
References 52

3. ADMISSION TO HIGHER EDUCATION IN THE UNITED STATES: THE ROLE OF THE EDUCATIONAL TESTING SERVICE Robert J. Solomon 53

Introduction 53
The Current Scene 55
How the College Board and ETS Serve the System 58
Conclusion 67

4. PUBLIC EXAMINATIONS IN AUSTRALIA John Philip Keeves 69

An Overview of Australian Secondary-School Examinations Systems 69
Special Issues of Certification and Selection 81

A Time of Change 85
References 87

5. EDUCATION IN SWEDEN: ASSESSMENT OF STUDENT ACHIEVEMENT AND SELECTION FOR HIGHER EDUCATION Sixten Marklund 89

Introduction 89
General Principles of Assessment 92
The Marking System 92
Standardized Tests 94
Non-standarized Tests 99
The Assessment Process 99
Free Choice of Study Route in School 101
Selection for Higher Education 101
Individual Results as Indicators of School Results 104

6. A BRIEF INTRODUCTION TO THE SYSTEM OF HIGHER SCHOOL ENROLLMENT EXAMINATIONS IN CHINA Lu Zhen 107

Historical Background 107
Examination Subjects 110
The Goals of the Examination 111
Grading the Examination 113
Admissions 114

7. DESIGNING THE ENGLISH LANGUAGE PROFICIENCY TEST IN CHINA Gui Shichun 115

Introduction 115
Stage One: Planning a Standardized Text 115
Stage Two: Pretesting 120
Stage Three: Administering the Test 128
Stage Four: Establishing the Norm 130
References 135

8. ASSESSING THE QUALITY OF EDUCATION OVER TIME: THE ROLE OF THE NATIONAL ASSESSMENT OF EDUCATIONAL PROGRESS (NAEP) Archie E. LaPointe 136

Testing in the United States 136
National Assessment of Educational Progress 141
The Assessment Process 143
Summary 146
Appendix: NAEP Report Card 147
References 150

9. CROSS-NATIONAL COMPARISONS IN EDUCATIONAL ACHIEVEMENT: THE ROLE OF THE INTERNATIONAL ASSOCIATION FOR THE EVALUATION OF EDUCATIONAL ACHIEVEMENT (IEA) John Philip Keeves 151

Introduction 151
Achievement in Reading 153
Achievement in Science at the Lower Secondary-School Level 153
The Effects of Time Spent on Learning 154
Opportunity to Learn in Mathematics and Science 155
Degree of Excellence 157
Retention Rates and Mathematics Achievement 157
Retention Rates and Science Achievement 159
Achievement in Developing Countries 159
Further Studies in Developing Countries 161
Further Cross-National Relationships 162
Conclusion 163
References 164

III. TESTING FOR THE IMPROVEMENT OF EDUCATIONAL MANAGEMENT 167

10. EXAMINATIONS AS AN INSTRUMENT TO IMPROVE PEDAGOGY Anthony Somerset 169

Introduction 169
Education and Examinations in Kenya 169
The Certificate of Primary Education Examination (CPE) 171
Instruments and Goals for Examination Reform 173
Some Effects of the Reform Program 187

IV. SUMMARY 195

11. IMPROVING UNIVERSITY SELECTION, EDUCATIONAL RESEARCH, AND EDUCATIONAL MANAGEMENT IN DEVELOPING COUNTRIES: THE ROLE OF EXAMINATIONS AND STANDARDIZED TESTING Stephen P. Heyneman 197

Common Background to Testing 197
Testing Issues 198
General Recommendations 206
Summary 216

V. SELECTED BIBLIOGRAPHY 217

TABLES

1.1. Budget of National Center for University Entrance Examination (NCUEE), 1983 15

1.2. Handicapped Test Applicants by Type, 1983-84 17

1.3. Subjects of Joint First-stage Achievement Test (JFSAT), 1985 18

1.4. Weighting JFSAT and Second-stage Examination 20

1.5. Selection Methods of Universities 21

1.6. Students Exempt from Taking JFSAT, 1985 22

1.7. Structure of NCUEE Research Division 22

3.1. Statistical Analysis of Test Takers 63

5.1. Diagnostic Tests and Achievement Tests in Sweden 97

5.2. Compulsory Written Tests in Swedish Upper-secondary Schools 98

5.3. Distribution of Students in Full-degree Programs, 1977 103

7.1. Uses of Different Types of Tests 117

7.2. Appraisal of Educational Objectives of English Proficiency Test (EPT) 118

7.3. Testing Syntax in Terms of Bloom's Educational Objectives 119

7.4. Distribution of Items in the English Proficiency Test (EPT) by Type and Complexity of Skill 120

7.5. Excerpt of EPT Item Analysis Printout 122

7.6. Computing Discrimination Index for Classroom Use 123

7.7. Biserial-correlation Coefficients of Two Models 124

7.8. Correlation Matrix of Different Sections Within EPT 125

7.9. Factor Loading after Varimex Rotation 126

7.10. EPT Items at Each Level of Difficulty 126
7.11. Tabulation of p-Value and r_{bis} Coefficients of EPT Test 127
7.12. Item Statistics of EPT's Listening Comprehension 128
7.13. Item Statistics of EPT as a Whole 130
7.14. Correlation Matrix Between Two EPT and Two TOEFL Tests 132
7.15. Test Evaluation Report Produced by GITEST II 132
7.16. Percentile Rank Distribution Table (Scholars to Study Abroad) 134
10.1. Standard Deviation of Kenyan District Mean Total Standard Scores, 1976-81 188
11.1. Summary Chart of Question Types 214
11.2 Kenya Primary Leaving Examination Test Items by Type and Year 215
11.3. China's EPT Items by Type, 1985 216

FIGURES

1.1. Organization of the Japanese Educational System Up to 1945 10
1.2. Organization of the Current Japanese Educational System 11
1.3. Structure of National Center for University Entrance Examination (NCUEE) 13
1.4. Configuration of NCUEE Computer System for Scoring 16
3.1. U.S. Students in Secondary and Post-Secondary Schools 56
3.2. Educational Testing Service (ETS) Corporate Structure 61
3.3 Steps in Developing a Test 66
5.1. School Structure in Sweden, 1984 91
5.2. Testing and Assessment in Swedish Schools 94
7.1. Flow Chart of Stages of Standardized Test 116

7.2. Positive and Negative Skewness 131

10.1. Mean Total Standard Scores for the CPE Examination, 1976-81 188

10.2. CPE Performance in Weighing Questions, 1979-81 194

PREFACE

Why should those concerned with the economics of developing countries pay attention to the problems of educational selection? In a competitive international environment, not choosing one's technical elite from among the brightest citizens can have a grave effect on economic performance. By one estimate, developing countries could improve their Gross National Product per capita by five per cent if they were to base leadership upon merit; by another estimate, the economic pay-off to developing countries would be three times more than the pay-off were OECD countries to reduce restrictions on third world exports.[1] Though the magnitude of effect may be debatable, the theory is reasonable.

The theory suggests that certain, but not all, elements of social selection are amenable to policy manipulation. Within the Education Sector there are basically three mechanisms--(1) whether a wide group of citizens enters school; (2) whether they stay in school; and (3) how the few of them chosen to attend university are selected. This discussion concentrates upon the use of tests to manipulate the third mechanism.

The use of selection tests concerns all developing countries, even those with near universal rates of attendance and grade-to-grade progression. When selection tests are used, educational systems are strongly affected. Selection tests tend to produce relatively tangible results by which to judge quality. Pressures felt on educational systems which use selection tests are sometimes broadly "popular" since, despite technicalities, results can be widely interpreted in the press and by various voluntary associations and interest groups. Selection testing draws attention at one specific time to a single, widely-understood indicator. It holds the school system accountable for results; and it fosters an open and continuing forum on the school system's ability to deliver results parallel to the public's expectations.

To be sure, selection tests create individual anxiety, which does, in fact, affect results. Moreover testing often highlights differences in the quality of educational inputs and learning opportunities among sub-groups of a population. Because of these problems, some progressive developing countries, such as Tanzania and Indonesia, began to dismantle their national systems of selection tests in the late 1960's and instead began to rely upon university-designed tests or upon selection criteria other than test results. The latter have been used to rectify past injustices to specific sub-groups; to insure fair geographical representation; and to recognize and reward abilities other than the academic. But both replacements have created unforeseen problems.

University tests have turned out to be logistically cumbersome as the number of universities expanded, and no more reliable as selection tests. Selection criteria other than testing have not turned out to be free of anxiety nor necessarily more fair. When school systems have relied upon particularistic criteria--political loyalty, family alumnae status, personal wealth, ethnicity, geography of birth--the effect has often been even more pernicious than the abandoned system of centralized testing. Those not selected by these criteria may feel that the choice was made unfairly. This has occurred, for instance, when the selection choice has been made subjectively on the basis of policital attitudes. Resentment and political backlash has also been known to occur when selection choices have been made on the basis of ethnicity. There the damage can be two-sided. Those selected may never feel as though they personally "deserved" the opportunity; those not selected may feel that the choice was made on an unacceptable basis since the rationale was to account for problems of group representation rather than individual worth.

---

[1] Sabastian Pinera and Marcelo Selowsky, "The Optimal Ability-Education Mix and the Misallocation of Resources within Education: Magnitude for Developing Countries." Journal of Development Economics 8 (1981): 111=31. Naheed Kirmani, Peirluigi Molajoni, and Thomas Mayer, "Effects of Increased Market Access on Exports of Developing Countries." IMF Staff Papers 31, No. 4 (1984): 661-84.

But not all tests are equally fair. Nor are all tests equally efficient. Today countries are shifting back towards 'the use of testing, and this, in turn, has raised new concerns. The result is the emphasis now being placed upon managing the details of selection testing better.

One example of this new emphasis on the need to purposefully manage the policies toward educational testing is that of China. In September 1984, the Chinese Government asked if the Economic Development Institute of the World Bank would be interested in assisting the officials in the Ministry of Education think through some of the policy options in the field of examinations and standardized testing. In response, in April 1985, a meeting was held in Beijing. Attending the meeting were all officials in charge of examinations at the provincial and national levels, technicians and psychometricians in charge of designing examination items, and senior university officials and planners in the Ministry of Education. Attending from outside the country were the chief executive officers of examination agencies in three OECD countries: from the United States, Robert Solomon (ETS); from Japan, Tadashi Hidano (National Center for University Examinations); and from the United Kingdom, John Reddaway (Cambridge University Examination Syndicate); directors of the National Assessment of Educational Progress and the International Association for the Evaluation of Educational Achievement; and experts on the examination systems in Sweden, Australia, and Kenya.

This book summarizes the descriptions of testing systems in selected OECD countries and the advice given to the government officials following the meeting in April 1985. The attention devoted to problems of logistics and to economies of scale are perhaps more pertinent to large, heterogeneous countries such as China; but the contention of this book is that many of the principles discussed at this meeting are applicable to developing countries generally.

These principles might be summarized as follows:

1. General Conclusion. No system of examinations is designed on technical grounds alone; each exists in a political environment. The system of aptitude testing in the United States exists because of the complex political prerogatives for communities to control their own curriculum. School-based assessments can function in Sweden because of the modest logistical prerequisites and consensus on criteria achievable in a small monoculture. Non-multiple choice formats can exist in Britan because the number of test-takers remains manageable and the definition of academic excellence has but modest variation from one university to another. Multiple testing by individual colleges can function in Japan because of the level of sophistication and motivation of the test-taking population. What this implies is that there is no OECD model which can be transferred to developing countries without forethought and adaptation. This forethought and adaptation apply to ex-colonial territories in the French and British commonwealths particularly.

2. Universities. The mechanism of selection will affect the quality of universities and therefore a nation's future. Since universities are expected to increase levels of self-financing and to be competitive internationally,

then universities should be given the responsibility of selecting their own students. Moreover, they should have school-based records of student accomplishments to assist them in their choice.

3. Test Agencies. Testing is a profession, highly susceptible to political interference. The quality of tests rest to a large extent on the ability of a testing agency agency to pursue professional ends autonomously. Agencies should therefore have their own source of finance from test fees. In the larger countries competition among agencies might be healthy. On the other hand, test agencies subsidized by the public sector should be expected to fulfill public functions. Among these functions should be to establish a strong system of analyzing test results and feeding this information back into the school system. Testing agencies should also share technical skills (item design, computer programming, etc.) and equipment with educational research functions of other agencies.

4. Tests. Where the test-taking population is high, geographically dispersed, culturally heterogeneous, or where the test employs a new national language, the test itself would benefit from a multiple choice format. Administrative and pedagogical effects would be maximized if questions were based upon curriculum; if they were open to *ex post facto* to public scrutiny; and if they were to include all levels of skill hierarchies, from recall to synthesis.

Stephen P. Heyneman

I

INTRODUCTION

Stephen P. Heyneman

and

Ingemar Fägerlind

Standardized tests were invented in China in the seventh century and later used for more than a thousand years as a means of selecting the nation's civil servants, judges, and military officer corps. The system built upon the candidate's ability to memorize, comprehend, and interpret classical texts. The system was abandoned in the early twentieth century and after 1949 replaced with a system of standardized academic achievement tests used primarily for selection to the nation's universities. During the Cultural Revolution this system was largely abandoned and replaced by such criteria as proletarian background and political activity. In the 1980s academic achievement has again become the most important determinant of university selection. The system has many problems and is deficient in terms of managerial efficiency, validity, reliability, and potential for research. The entire process of design, administration, and grading, in which 29,000 teachers are involved, takes up to three months. With more and more students applying for higher education, the Chinese authorities planned to reform the system of admission to higher education. The seminar on the uses of standardized tests and selection examinations, held in Beijing in April 1985, was one of the steps taken as part of the decision-making process.

The Importance of Education to Economic Growth

The recent Chinese economic reform, the massive modernization program, and the huge expansion in the field of education are among the most impressive attempts in human history to move a developing nation into an industrial one. In May 1985 the Central Committee of the Communist Party of China (CPC) made very important decisions about educational reforms. "Reform of China's Educational Structure--Decision of the CPC Central Committee," published in Beijing in May 1985, views education as necessary for economic growth.

The beginning of that document outlines the enormous task that the Chinese nation has in front of it:

> We need to train millions upon millions of workers in industry, agriculture, commerce, and other fields, who are well educated, technically skilled, and professionally competent. We also need to train tens of millions of factory directors, managers, engineers, agronomists, economic experts, accountants, statisticians, and other economic and technological personnel who are equipped with modern knowledge of science, technology, and economic management and imbued with a pioneering spirit. And we need to train tens of millions of educators; scientists; medical workers; journalists; editors; publishers; and workers in the field of law, foreign affairs, and military affairs, as well as Party and government workers who can keep abreast of developments in modern science and the technological revolution.

The effectiveness of the modernization program depends upon the quality and quantity of the skilled manpower available. At the same time it is stressed that the development of education must take place within the framework of limited financial and material resources. Education must become more effective, even if central- and local-government appropriations for educational purposes increase at a rate faster than the increase in the state's regular revenues and even if the average expenditures on education per student also increase steadily. Because the seminar was held at the same time that the overall reform plan in education was being drafted, it must be looked upon as one example of how Chinese authorities want to be informed about how problems are solved in other countries before they make their own decisions. It is totally clear, however, that the Chinese plan to have a more efficient selection system and that they want to develop different types of tests that have good predictive validity both for academic and professional success.

The Importance of Testing in Education

Educational testing is not only a scientific endeavor, it is also a political act. What kind of knowledge is important depends upon the goals of the society. Testing is a way of judging quality, and it is usually important for politicians, policymakers, taxpayers, and parents to know the quality of educational outcomes. If testing is done repeatedly, it is also possible to assess the quality of education over time. It is important to know if the quality of educational outcomes is going up or down over time in a country, a province, in different parts of the educational system, or in a certain school. Measuring educational outcomes over time requires special tests with "anchor" items used in each test given at regular intervals.

Many educationists are interested in the quality of education in their own country compared to that in other countries. Since the beginning of the 1960s the International Association for the Evaluation of Educational Achievement (IEA), a non-governmental organization made up of professional institutes of education, has tried to explain differences in results between countries. Studies comparing performance on achievement tests, attitudes towards education, and the proportion of students in an age-cohort that perform above a certain level have been conducted in some forty countries all over the world.

Testing can also be used to improve classroom pedagogy. Tests can show where additional teaching and learning must take place. Means, standard deviations, and item analyses can be used as raw material for analyzing why errors occur in learning patterns. After analyzing these patterns, teaching methods could be suggested on how to overcome errors, and negative results could be converted into encouragement and specific suggestions for improvement.

The most common use of testing in education is for selection purposes. Today testing is used in most countries of the world for choosing students who qualify for further training and for later societal leadership. Selection examinations may be designed so that they emphasize different criteria. Tests could be designed to select innovative and highly-specialized or hard-working, well-prepared students. They can also select students according to general intelligence or specific skills. It is of utmost importance that tests be designed that are valid in determining future success both in studies and work.

The Importance of Improving the Quality of Testing

Essay tests have a long history both in China and in other countries. They have some advantages, for example, when measuring what overview a student has in a certain field. However, such tests also have many disadvantages. Grading is time-consuming and often unreliable. For this reason multiple-choice questions are becoming more popular in selection tests. Different ways of improving the quality of testing were discussed at the seminar by experts from many institutions around the world.

The Organization of the Seminar

The seminar was planned more than one year in advance, and the purpose was spelled out in the following way:

(1) To bring to China the best experience on the uses and functions of standardized tests and examinations.

(2) To familiarize Chinese authorities with costs, benefits, and managerial requirements of various national systems of tests and examinations.

(3) To allow Chinese authorities to discuss and question firsthand the pros and cons of different national systems.

(4) To help the Chinese authorities choose among various alternatives in making improvements.

About one hundred persons, most of whom had been involved in the construction of the present entrance examinations, participated in the seminar. Some sessions were attended by up to two hundred fifty persons. The seminar started April 11, 1985, and went on until April 25. The Foreign Language University of Beijing provided accommodations as well as simultaneous translators from their group of students in training to be United Nations translators. Lectures were usually followed by group discussions in which the foreign visitors were part of the different groups. All the papers used at the seminar had been translated into Chinese and were handed out the day before the lecture. As the participants were enthusiastic about the group discussions, some changes were made in the schedule to accommodate more group work. At the end of the seminar written evaluation forms were filled in by the participants.

Potential Uses of This Book

As the topics discussed at the Beijing seminar are important issues in most countries of the world, it is our hope that this publication can be useful at regional and/or national seminars on the same topic. It can also be used in courses on comparative education and development at the university level and in many schools of education.

II

TESTING FOR SELECTION TO UNIVERSITY

ADMISSION TO HIGHER EDUCATION IN JAPAN

Tadashi Hidano

The History of Examinations for Admission to Higher Education

Examinations before World War II

The Japanese examination system to select officials was introduced from China. In 701 A.D. the imperial examination system called Kuko no Sei was established, and talented people were selected according to their results on written examinations. This system did not play an important role in aristocratic society, however, and lasted only a few centuries.

Japan's modern education system was established in 1872, after the models of European and American education systems. It was organized in three progressive stages of elementary, secondary, and higher education. The basic structure of education up to 1945 is illustrated in figure 1.1.

Higher schools and preparatory courses for universities, colleges, and higher normal schools selected their entrants from graduates of middle schools and girls' high schools.

Each institution offered its own examination. The essay-type written examination was most frequently used for selection. The teaching staff of each institution was responsible for writing questions, scoring answer sheets, and determining successful applicants. In contrast to qualification examinations, enrollment quotas were decided and announced in advance, and applicants were ranked according to their examination scores. Those applicants with ranking within the quota were admitted.

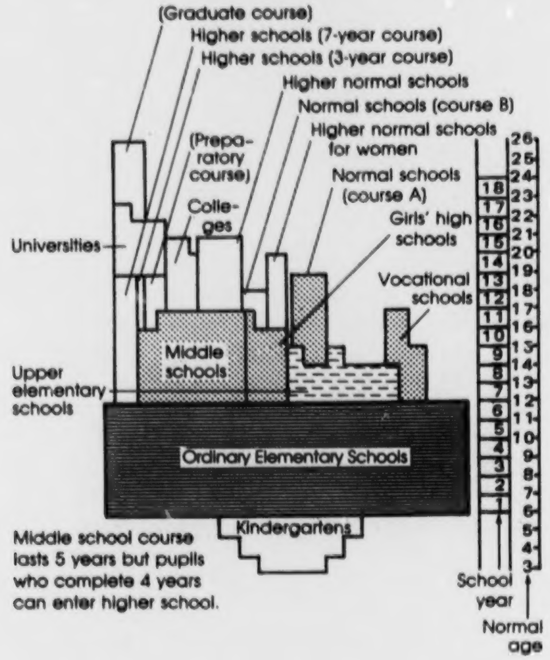
Though university selection procedures have traditionally been carried out by each university independently, a unified admission system was attempted twice before the World War II. In 1902 and 1917 all higher schools had common examinations. Successful entrants were distributed according to their ranking in the score and their order of preference. The unified examination system was abolished after a few years.

Higher Education after World War II

After World War II, the Allied Occupation attempted a radical reform of the education system. The "multi-track" system of school education was changed into a "single-track" system, and almost all schools became coeducational. Figure 1.2 illustrates the organization of the present school system.

Since the reform plan was put into effect in 1947, numerous colleges and universities have been established. As of 1984 there were 460 universities (95 national, 34 local public, and 331 private), 448 junior colleges, and 62 technical colleges. Total enrollment in universities was 1,734,080, and in junior colleges, 375,450. A total of 21 percent attended national universities; 3 percent, local universities; and 76 percent, private universities.

Figure 1.1. Organization of the Japanese Educational System Up to 1945.

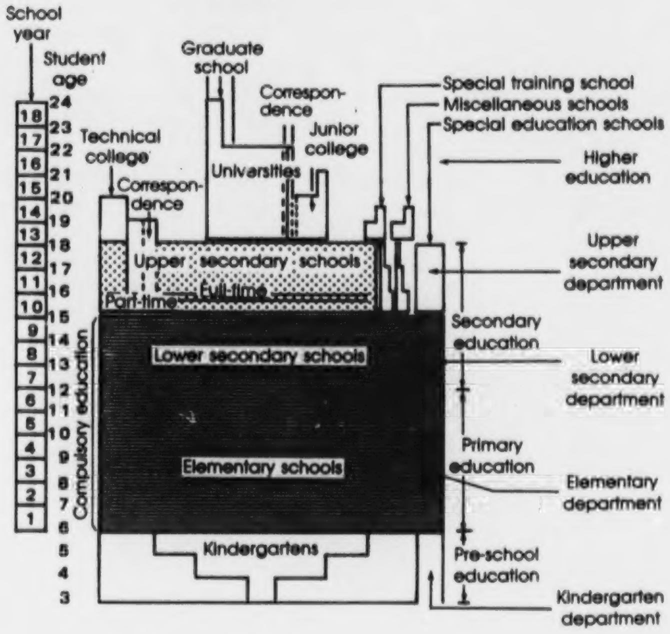

(Graduate course)
Higher schools (7-year course)
Higher schools (3-year course)
Higher normal schools
Normal schools (course B)
Higher normal schools for women
(Preparatory course)
Colleges
Normal schools (course A)
Girls' high schools
Universities
Vocational schools
Middle schools
Upper elementary schools
Ordinary Elementary Schools
Kindergartens
School year
Normal age
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26

Middle school course lasts 5 years but pupils who complete 4 years can enter higher school.

The reform encouraged pupils and their parents to aim for higher education as a means of upward social mobility. The proportion of the age group going to upper-secondary schools, which are not compulsory, steadily increased and reached 94.3 percent in 1981. The proportion of the age group going to universities and junior colleges, which had been less than 5 percent before World War II, rapidly increased after the war and reached 38.6 percent in 1976. These increases, however, hit the ceiling and have been sluggish in the past few years.

Since the capacity of universities has not expanded more rapidly than the increase in applicants, competition to gain entry is always high. Moreover, differences exist among universities in terms of tradition, professional fields, opportunities for employment after graduation, location, and quality of students. The differences are reflected in prestige. This had led to intense competition among those trying to enter higher-ranked universities.

Figure 1.2. Organization of the Current Japanese School System


School year
Student age
24
23
22
21
20
19
18
17
16
15
14
13
12
11
10
9
8
7
6
5
4
3
18
17
16
15
14
13
12
11
10
9
8
7
6
5
4
3
2
1
Graduate school
Correspondence
Junior college
Technical college
Correspondence
Universities
Special training school
Miscellaneous schools
Special education schools
Higher education
Upper secondary schools
Full-time
Part-time
Upper secondary department
Secondary education
Compulsory education
Lower secondary schools
Lower secondary department
Primary education
Elementary schools
Elementary department
Kindergartens
Pre-school education
Kindergarten department

Examinations after World War II

After World War II three plans to improve the selection system of higher education were tried. In 1947 the scholastic aptitude test was introduced to the admission systems of Japanese universities. The test was constructed by the National Institute for Educational Research, a branch of the Ministry of Education. All national universities, and later all universities including local public and private ones, used the test scores for selecting entrants. The test system was abolished in 1954.

In 1963 a nonprofit institution, called the Talent Development Institute, was established. It offered three kinds of tests: the scholastic aptitude, the vocational aptitude, and the achievement test. Though many upper-secondary schools used these scores for vocational guidance, only a few universities used the test score to select entrants. The institute was dissolved in 1968.

In 1971 the National University Association organized a special committee to investigate improvements in the selection system. In 1976, after concentrated investigation and three experimental tests, the association decided on a two-stage test, in which all national universities administer the same achievement test jointly at the first stage, and each university offers its own selection examination at the second stage. The same year the Local Public University Association decided to join the new system.

The Joint First-stage Achievement Test (JFSAT) aims at evaluating applicants' attainment in basic and general studies at upper-secondary schools. The second-stage examination offered by individual universities assesses the abilities and aptitudes of applicants which the university requires of its entrants.

The National Center for University Entrance Examination (NCUEE) was established in May 1977 as the institute in charge of constructing and administrating the JFSAT in cooperation with national and local public universities. The center also conducts extensive research on improving the system and methods of selection. In January 1979 the first JFSAT was administered, and one week later a supplementary examination was conducted.

The National Center for University Entrance Examination (NCUEE)

Organization and Administration of NCUEE

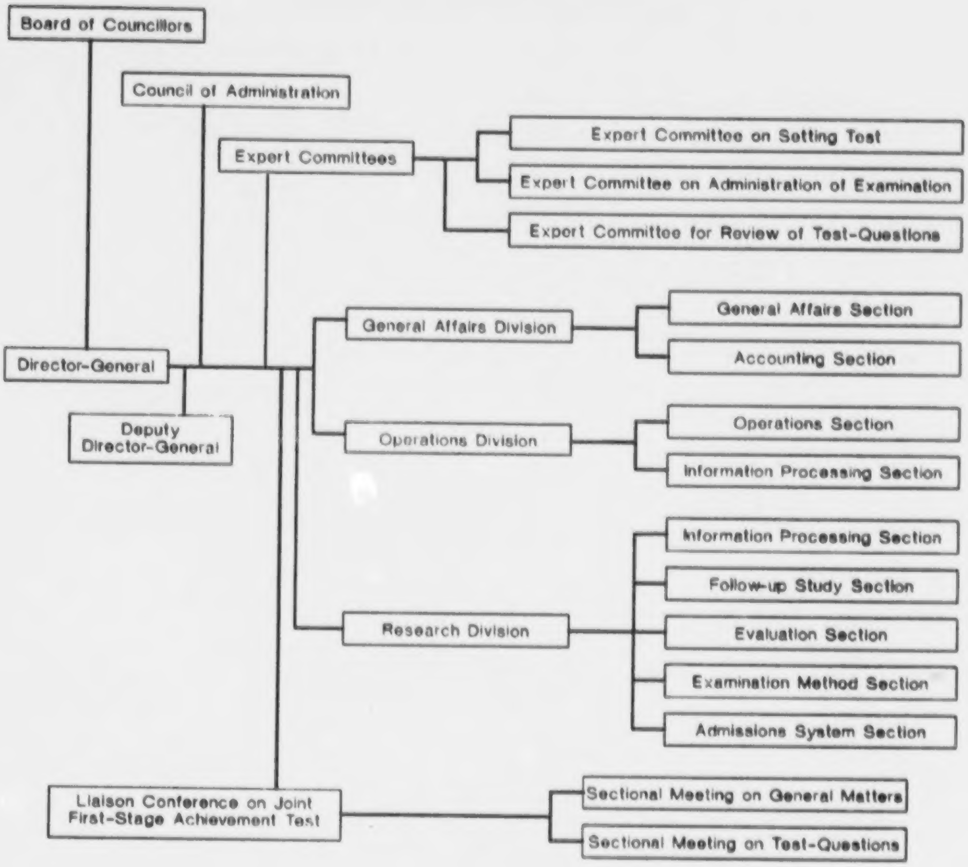
The National Center for University Entrance Examination (NCUEE) has a permanent staff of eighty-eight as of January 31, 1985. The structure of the center is shown in Figure 1.3.

The Board of Councillors consists of fifteen members who are presidents of national universities and persons of learning and experience. The board discusses the projects of the center and other important administrative matters and gives advice to the director-general.

The Council of Administration is made up of twenty-one members who are professors from national universities, persons of learning and experience, and professors from the center. The council discusses practical matters related to JFSAT and the administration of the center, upon the request of the director-general.

The Expert Committees consist of faculty members of national universities. The committees are the Expert Committee on Setting the Test, which prepares questions for the JFSAT. The Expert Committee on Administrating the Examination consists of twenty members who plan and attend to administrative matters related to the JFSAT. The Expert Committee for the Review of Test Questions inspects and checks the content and construction of the test questions set by the Expert Committee on Setting the Test.

Figure 1.3. Structure of National Center for University Entrance Examination (NCUEE)


Board of Councillors
Council of Administration
Expert Committees
Expert Committee on Setting Test
Expert Committee on Administration of Examination
Expert Committee for Review of Test-Questions
Director-General
Deputy Director-General
General Affairs Division
General Affairs Section
Accounting Section
Operations Division
Operations Section
Information Processing Section
Research Division
Information Processing Section
Follow-up Study Section
Evaluation Section
Examination Method Section
Admissions System Section
Liaison Conference on Joint First-Stage Achievement Test
Sectional Meeting on General Matters
Sectional Meeting on Test-Questions

The Liaison Conference on JFSAT keeps in contact with upper-secondary schools so that their opinions and requests affect various aspects of the JFSAT.

The Sectional Meeting on General Matters consists of twenty representatives from upper-secondary schools, boards of education, and the center. They discuss and contact upper-secondary schools on general matters related to the JFSAT.

The Sectional Meeting on Test Questions consists of sixty-six representatives from upper-secondary schools and local boards of education recommended by the boards of education of the metropolis and districts (three members for each subject) and persons in charge of setting test questions for each subject of the center.

All activities of the center are supported by the government, and the examination fees (9,000 yen per capita) belong to the government.

Time Schedule for Selecting National and Local Public University Entrants

National and local public universities announce the basics of the second-stage entrance examination by the end of July. Each association, national and local public universities, and the NCUEE publish a guidebook in September. Applications for the JFSAT are accepted from November 1 to 10.

National and local public universities release details of the second-stage entrance examination. Application forms and other documents are distributed to applicants by December 25. The number of applicants to faculties in national and local public universities is published early in January. The JFSAT is administered on the last or second last Saturday and Sunday of January. The supplementary examination is administered on the next Saturday and Sunday to absentees of the regular examination. General results of the JFSAT are published in February. National and local public universities accept entrance applications from February 9 to 15. The universities that employ the two-stage selection examination system or admissions by recommendation announce results by February 26. All national universities and the majority of local public universities start the second-stage entrance examination on March 4. The remainder of local public universities administer the second-stage entrance examination on and after March 5. National and local public universities announce successful applicants by March 20. Some national and local public universities carry out supplementary selection on and after March 21.

Setting Test Questions

Questions of the JFSAT are set by the Expert Committee on Setting the Test, which is divided into working groups according to each examination subject.

Administrating the Examination

The JFSAT is administered to more than 330,000 applicants at 280 examination halls throughout the country. Each hall is linked to the center by means of a communication system consisting of facsimile and telephone lines. The total number of examination rooms is about 5,640. The same test questions are administered at the same time. All national and local public universities join in administering the JFSAT. About 42,500 staff cooperate in the administration.

Scoring and Reporting of Scores

All marked sheets are collected at the center, read by an optical mark reader system, and scored by the computer.

The test scores of the individual applicant are reported to the university which received his or her application. The applicant is not informed of his or her scores from the center.

Scores cannot be sent to more than one national university, but students may apply in addition to some local public universities which administer their second-stage examinations on and after March 5.

Publishing Results

The questions, the keys (correct answers), and the weights to questions are disclosed by the center just after the administration of the test. A brief summary of the test results is published prior to the application deadline for the second-stage examination. These processes enable the applicants to estimate their own scores and rank in order to chose an optimum university and faculty for the second-stage examination.

Table 1.1. Budget of National Center for University Entrance Examination (NCUEE), 1983

Item	Amount (thousand yen)
Personnel Expenses	344,712
General/Operational Expenses	1,073,707
JFSAT Implementation Expenses	1,367,710
JFSAT Implementation Expenses Distributed to Universities	923,030
Facilities Improvement Expenses	515
Total	3,709,674

Special Arrangement for Handicapped Applicants

Based on requests from handicapped applicants, the center provides special measures according to the nature and degree of the disability. For example, it may prepare test papers in Braille or with enlarged letters, extend test hours, set up special examination rooms, permit an interpreter of sign language or other helpers to attend, or permit answering questions in written script instead of coloring in letters on a mark sheet.

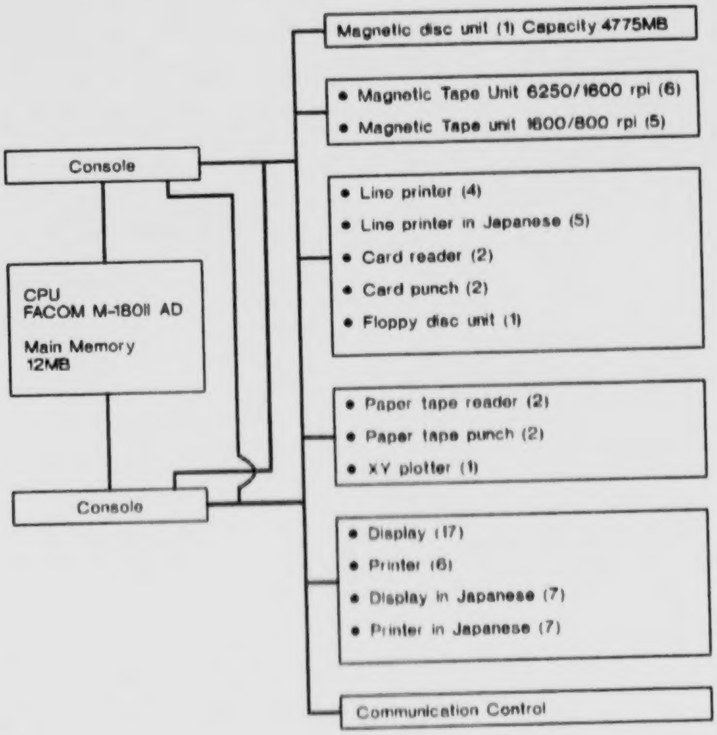
Public Relations

"The Annual Bulletin of the National Center for University Entrance Examinations" is a report on the outline of the JFSAT and the activities of the center.

"Opinions and Evaluations of Test Questions Set for the Joint First-stage Achievement Test" is an annual report compiling opinions and evaluations of the JFSAT questions submitted by high schools and educational research groups; it also includes the center's opinions on the submissions.

"The University Entrance Examination Forum" is a periodic information magazine which reports, explains, and introduces research activities focusing on the further improvement of the selection methods for university entrants and the actual state of the implementation of the JFSAT.

Figure 1.4. Configuration of NCUEE Computer System for Scoring


Magnetic disc unit (1) Capacity 4775MB
• Magnetic Tape Unit 6250/1600 rpi (6)
• Magnetic Tape unit 1600/800 rpi (5)
Console
• Line printer (4)
• Line printer in Japanese (5)
• Card reader (2)
• Card punch (2)
• Floppy disc unit (1)
CPU
FACOM M-180II AD
Main Memory
12MB
• Paper tape reader (2)
• Paper tape punch (2)
• XY plotter (1)
Console
• Display (17)
• Printer (6)
• Display in Japanese (7)
• Printer in Japanese (7)
Communication Control

Optical mark reading systems are listed below.

W2300 system(6)		W201 system (1)	1
CPU (with a 16 KW main memory)		CPU (with a 32-KW main memory)	1
W301 scanner (made in U.S.)	6	W201 scanner (made in U.S.)	
Reflected-light detecting system		Reflected-light detecting system	
Maximum-reading speed:		Maximum-reading speed:	
18,000 sheets/hr		10,200 sheets/hr	
Magnetic-tape unit	12	Magnetic-tape unit	1
Display	6	Display (with a magnetic cassette tape unit	
Paper tape reader	1	Line printer	1

Table 1.2. Handicapped Test Applicants by Type, 1983-84

Type	1984	1983
Blind	42	54
Deaf	70	69
Other handicapped	80	65
Total	192	188

"The Joint First-stage Achievement Test" is public-relations material which includes explanations of the present status and tasks of the JFSAT, a report on the test results, and an outline of entrance examinations for national and local public universities.

"The Guide to National and Local Public Universities" introduces the characteristics and history of each university and includes an outline of its second-stage entrance examination. It is compiled jointly by the National University Association, the Local Public University Association, and the NCUEE.

Setting and Reviewing the Joint First-Stage Achievement Test (JFSAT)

Committee for Setting Questions

The Expert Committee on Setting the Test consists of about 230 members selected from professors and associate professors at national universities throughout Japan. Each working group, which consists of twelve to fifteen members, holds about ten three-day sessions a year to set several forms of questions. Each member's term is two years, with one-half of the members selected each year.

Process of Setting Questions

Every spring each working group begins constructing new questions. After a half year it completes a draft of the questions, which is reviewed by the Expert Committee for the Review of Test Questions. Also, a joint meeting of working groups in the same subject area is held to adjust the content of questions and levels of difficulty among different subjects. After a final proofreading, the test questions are printed.

Table 1.3. Subjects of Joint First-stage Achievement Test (JFSAT), 1985

Subject area	Subject	Subject to be chosen
Japanese Language	Japanese Language	
Social Studies	Modern Society, Ethics, Political Science, and Economics Japanese History World History Geography	Two subjects must be chosen: one from Modern Society, Ethics, Political Science, or Economics; and the other one from Japanese History, World History, or Geography.
Mathematics	Math. I Math. II Engineering Math. Bookkeeping and Accounting	All questions on Math. I must be answered. One subject from Math. II, Engineering Math., and Bookkeeping and Accounting must be chosen.
Science	Science I Physics Chemistry Biology Geophysics-Astronomy	Two subjects, Science I and one from Physics, Chemistry, Biology, or Geophysics-Astronomy must be chosen.
Foreign Languages	English German French	One subject from English, German, or French must be chosen.

Owing to the revision of upper-secondary-school curriculum, the subjects changed in 1985.

Principles for Setting Questions

The general principles for setting questions are as follows:

Questions should be set in accordance with the upper-secondary-school course of study and textbooks.

In all subjects, uniformity in style and level of questions is desired.

The mean score for each subject should be more than 60 percent of the full mark.

Research on the result of past examinations should be reviewed as a guide for setting questions.

At every stage of question setting a thorough discussion among members is essential.

The security of confidential materials and information is essential.

Tasks and Responsibilities of Committee Members

Each working group of the committee is responsible to write objective questions. Multiple choice is used in almost all questions, but in mathematics the short answer is used.

The group sets several forms which are expected to be as parallel as possible. One of them is administered at the regular examination, and another one, at the supplementary examination.

All members of the committee are responsible for the security of confidential materials and information.

The committee members' travel expenses and compensation for services are paid by the center.

Review by the Expert Committee

The committee is divided into sub-groups according to each subject. After reviewing the draft questions and correct answers (keys), each group suggests improvements. Then the working group of the Expert Committee on Setting the Test modifies the draft, taking the suggestions into consideration.

Printing and Distribution

After the final proofreading, test questions are printed and stocked. They are distributed to examination halls before the administration of examination. Only answer sheets are collected at the center.

Scoring

All answer sheets are read by an optical mark reader system and scored by computer. Correct answers (keys) and the weight of each item are decided beforehand by the Expert Committee on Setting the Test.

Analysis of Test Scores

In each subject, all answer sheets are classified into five groups according to their total score. The response rate for each choice for each item is calculated for each group.

The difficulty of each discrimination power is estimated from the diagram which represents the response rates of all choices for each item. The correlation coefficients of scores between different item groups and between subjects are calculated and many other statistical analyses are carried out.

Critical Review of Test Questions

After the administration of the JFSAT, a Sectional Meeting on Test Questions is held in February and March. Teachers of upper-secondary schools hold discussions with representative members of the Expert Committee on Setting the Test. They write a critical review of test questions regarding content, difficulty, and format. In addition, representative national associations of education in each subject area are asked for their evaluation of the test questions.

Storing Information about Test Questions

All information about JFSAT results are stored. Also, as many as possible of the questions which are administered at entrance examinations of national, local public, and private universities are collected.

Table 1.4. Weighting JFSAT and Second-stage Examination

	Number of Univ.	Univ. disclosing weight	Univ. giving heavier weight to JFSAT	Univ. giving equal weight	Univ. giving heavier weight to second-stage exam	Univ. faculties using different weighting	Univ. giving different weights to subject areas
National	94	85	39	9	7	30	56
Local Public	34	30	17	0	3	10	21
Private	1	1	0	0	1	0	1
Total	129	116	56	9	11	40	78

University Selection Procedure

Regular Selection Procedure

Each university determines successful candidates among its applicants. National and local public universities comprehensively consider all materials: the JFSAT scores, scores of their second-stage examinations, and credentials from upper-secondary schools. No private universities, with one exception, use the JFSAT score because they have not joined the system.

Among national and local public universities fifty-six give heavier weight to the JFSAT score, while eleven give heavier weight to their own second-stage examination results. Many universities use a differentially weighted sum of scores of subject areas instead of an equally weighted sum of scores.

Most universities offer achievement examinations, but the examinations are limited to a few subjects which are needed for the entrants to study at the faculty or department to which they apply.

Applicants who failed the second-stage examinations can apply to a few national or local public universities which admit a limited number by supplementary selection. This is carried out on and after March 21.

Table 1.5. Selection Methods of Universities

	National Univ.	Local Public Univ.	Private Univ.
Univ. exempting achievement examination	4	4	-
Univ. offering achievement examination	90	30	1
Univ. offering composition	60	19	1
Univ. interviewing applicants	39	9	1
Univ. testing practical skills	55	5	-

At the second-stage examination, diverse selection methods are used.

(As of 1985)

Special Selection Procedure

Some universities admits applicant on the bases of recommendations from the upper-secondary schools. Half of them require that applicants take the JFSAT, and these scores are used for selection. The other half exempt them from the JFSAT.

Universities which offer a special examination to Japanese students who have completed a secondary-school course overseas exempt them from taking the JFSAT. Adult students and foreign students are also exempt from taking the JFSAT at universities which offer a special examination for them.

Table 1.6. Students Exempt from Taking JFSAT, 1985

		National Univ.	Local Public Univ.	Private Univ.
Selection based on recommendation	JFSAT required	40	7	----
	JFSAT exempted	37	11	----
Special selection for Japanese students overseas		21	5	----
Special selection for adult students		11	1	----
Supplementary selection		31	1	1

Research Activity in the National Universities

In each national university, a research committee has been set up to study the selection method and ways to improve it. In 1980 the Conference on the Admission Research of National Universities was organized. This organization consists of the research committees of all national universities and the research division of the center. It holds an annual meeting and publishes a research bulletin every year. It also promotes joint research activities among member institutions.

Research

Table 1.7. Structure of NCUEE Research Division

Section	Prof.	Guest Prof.	Associate Prof.	Assistant	Main Theme
Information Processing	1	-	1	1	1) Improvement of information-processing system used in the university entrance examination 2) Statistical analysis of JFSAT questions. 3) Determination of optimal weight of items.

Table 1.7. Structure of NCUEE Research Division (Continued)

Section	Prof.	Guest Prof.	Associate Prof.	Assistant	Main Theme
					4) Analysis of applicant's choice of universities 5) Local differences in JFSAT scores.
Follow-up Study	1	-	1	1	1) Interrelation of the scores of JFSAT and second-stage examination. 2) Follow-up study on students' activity and academic record in the university.
Evaluation	1	1	1	1	1) Evaluation of JFSAT questions and second-stage examination questions. 2) Analysis of test scores. 3) Validity and reliability of objective test. 4) Distribution of scores and scaling. 5) Method of writing questions in Braille.
Examination Method	1	1	1	1	1) Improvement of examination methods. 2) Test questions and upper-secondary-school curriculum. 3) Comparison of Japanese examination methods with those of other countries.
Admission System	1	-	-	1	1) Research on selection system of university. 2) Comparative study on selection system in Japan and other countries. 3) Historical study on admission systems.
		Joint research			1) Multivariate analysis on JFSAT scores. 2) Relationship between university entrance examination and high school and college education.

Technical and Social Issues

Academic Achievement vs. Aptitude

In the modern educational system, institutes of higher education employ mainly essay-type examinations of academic achievement for their selection of entrants. Though oral examinations were used in some cases, the main selection device was written examinations.

Scholastic aptitude tests were administered only from 1947 to 1954 and from 1963 to 1968. These tests were composed mainly of two sections: verbal and mathematical. The scholastic aptitude test was not used as the main selection device because the correlation of its scores to academic records at the university level was lower than achievement examination scores. Also, achievement examinations were a more reliable index of academic attainment in the upper-secondary schools.

Essay Examination vs. Objective Test

In Japan as in other countries, the essay examination has a long tradition, while the objective, machine-readable test was introduced after World War II. In elementary and secondary schools objective tests have been frequently used for classroom teaching as well as for selection, but many faculty members at universities are still critical of them. Though the efficiency, objectivity, reliability, and validity of the JFSAT have been proved by many studies, the issue is still highly controversial in Japan.

Uniform Requirements vs. Requirement by Each University or Faculty

The JFSAT covers five subject areas. In social studies one can choose one subject from Japanese history, world history, and geography. In mathematics one can choose one subject from mathematics II, engineering mathematics, and bookkeeping-accounting. In natural science one can choose one subject from physics, chemistry, biology, and geophysics-astronomy. In foreign languages one can choose one subject from English, German and French.

If the JFSAT is the minimum requirement for admission to national and local public universities, the requirement should be uniform. But if diversity of selection is more important, different requirements by each university or faculty should be allowed. This is a hotly debated issue today.

Difference of Mean Scores Between Subjects in the Same Subject Area

The JFSAT is not a standardized test in its strict meaning. It is unable to try out a sample test before the test is administered because the security of the JFSAT is most important. All questions are disclosed just after they are administered. It is unable to equate scales between different subjects. Also raw scores rather than standardized scores are used. In these circumstances it is unavoidable that mean scores differ between subjects in the same subject area.

Some studies done by the research division of the center show that the difference of mean scores between subjects reflects the difference in scholastic ability of applicant groups who have chosen the subject rather than the difference of the level of difficulty of the questions.

Multiple Opportunities for Taking the Examination

Before 1973 national universities formed two groups. The universities in one group administered their entrance examinations on and after March 3, and universities in the other group, on and after March 23. Thus, applicants were able take the examinations for two national universities. Since 1979 all national and almost all local public universities administer their second-stage examination on the same days. In this new system applicants can take the examination for only one university a year. The pros and cons of the old and new systems are being discussed now in the National University Association.

Applicants' Choice of University

After taking the JFSAT, many applicants send their estimated scores to some private guidance institutions. These institutions make forecasting tables of success for each university and faculty. Guidance teachers at upper-secondary schools use these tables for advising applicants on their choice of university and college.

Guidance teachers and applicants depend so heavily on the forecasting tables that they scarcely consider the applicants' own ability and aptitude other than academic achievement. Faculty members in the university complain that their entrants are much too homogeneous in terms of the JFSAT scores. However, research results do not support their complaint.

The JFSAT system is also criticized because the ranking of universities and faculties became more explicit after 1979 when the JFSAT was introduced. Uniform requirements, the single opportunity of taking the examination, and ranking universities are interrelated issues which are very difficult to solve.

EXAMINATIONS FOR UNIVERSITY SELECTION IN ENGLAND

John Lewis Reddaway

Historical Origin and Functions

The examinations which are the mainstay of university selection in England have their origins in the development of university-regulated examinations for schools in the late 1850s. Examinations and inspections for schools were developed by the universities of Oxford, Cambridge, London, and Durham in response to requests from the schools for them to provide these services and not as a result of government policy. However, the establishment of the School Certificate examination in 1917 by the government brought into being a nationally recognized academic school examination, albeit a pluralistic system of equivalent examinations offered by the different universities. The School Certificate examination was a "group" examination requiring success in one subject from each of five groups, namely English, languages, science and mathematics, music, and manual subjects. The examination was designed primarily as a test of achievement at age fifteen to sixteen, but the achievement of credits in all subjects led to exemption from the Matriculation Certificate examinations, signifying that the holder was qualified to enter university. This was extended in 1922 by the introduction of the Higher School Certificate, a more specialized examination based on two years of work in the school sixth form for seventeen-to-eighteen-year-old pupils. Eventually the Higher Certificate came to be the more significant with regard to university selection. Increasingly it was argued that the group examination had unfortunate byproducts. Not all pupils sought entry to university or were suited to work across the full range of subjects, but all were forced into the same curricular mold by the inflexibility of the examinations. In 1951 the General Certificate of Education (GCE) was introduced at two levels: ordinary level served the sixteen-plus age group, while advanced level replaced the Higher School Certificate as an examination, designed for those two years older. These were single-subject examinations, and candidates could present themselves in as many subjects as they or their teachers saw fit. As independent bodies the universities then met and agreed to define matriculation requirements in terms of the GCE. Generally they asked for passes in five or six academic subjects, at least two of them at the advanced level, as a minimum requirement. Faculties often imposed additional requirements for passes in specified subjects. As demand for university places began to outpace supply, a system of gradings to differentiate those above the pass mark at the advanced level was introduced to facilitate the use of the examination as a selection device.

Thus, from this point onward the GCE examinations performed various functions, including the three following:

(1) First and foremost they assessed pupils' success in the courses they followed at school. As achievement tests they therefore measured

ability moderated by, among other factors, effort and instruction. It is against this criterion alone that the examiners would wish to be judged. Other functions are subsidiary and flow from this.

(2) They came to function as a kind of qualifying examination, a minimum general competency measure, through widespread adoption of the erstwhile matriculation concept by many professions and employers. Thus, for many occupations it is necessary to pass in five subjects at the GCE ordinary level prior to entry into employment. Other professions use the old matriculation requirements of five subjects, two at advanced level, as their requirement for recruits. Of course, meeting such requirements does not guarantee a job. It merely qualifies one to apply for one, thus acting as a preliminary hurdle before the real process of selection begins. This has spared schools the necessity of providing a great diversity of courses aimed at preparation for entry into specific occupations and provides employers and others with a ready, if crude, classification of abilities.

(3) They form part of the mechanism for competitive selection for higher education (and employment too). Other factors may be equally important, however, depending on circumstances. This will be discussed in more detail later.

Developments since the introduction of the GCE in 1951 have not changed this position. The most notable of these was the introduction of the Certificate of Secondary Education (CSE) in 1965 as a complementary examination to the ordinary level of the GCE, designed for pupils in the 40th to the 75th percentile range of ability, and thus below the GCE's target range. At the present time the fusion of the GCE ordinary level and the CSE into a single system of examinations for the sixteen-plus age group is being planned, but the advanced level, the principal university selection device, will be unaffected.

Organization of the General Certificate of Education (GCE)

The GCE is provided by seven English GCE boards. These are independent bodies, all but one being connected with universities. Each operates nationally, schools being free to choose any examination from any board.

Many schools do exercise this option and approach different boards for their examinations in various academic subjects. Thus, there is no regional system. Rather, there is a pluralistic national one. National coordination is provided through the Secondary Examinations Council, which appraises and approves all A-level syllabuses and monitors the boards' work. The various boards collaborate from time to time in exercises designed to look at the comparability of their examinations and their grading standards so as to coordinate them insofar as it is possible (Bardell, Forrest, and Shoesmith, 1978).

A pluralistic system of this sort is desirable in the British context in order to maximize the choice available to individual schools, so enhancing their freedom to teach the curriculum which they feel best suits their pupils' needs. In a small, densely populated country it functions effectively. In a less densely populated state regionalization might be more attractive, especially if schools have less choice of curriculum than is the case in England. While our system may appear administratively untidy, it facilitates the close correspondence between the courses taught in the schools and examinations, which is at the root of the validity of the latter and a much prized feature. In addition, the diversity within the system leads to examinations which involve relatively small numbers of candidates, and so enables senior examiners to remain in close contact with the process of marking and standards fixing. At the same time it promotes the continuity of expertise so essential to consistency in a system which does not shrink from the problems inherent in making human judgments.

My own board, the University of Cambridge Local Examinations Syndicate (UCLES), will serve to illustrate the way in which GCE boards work, although it is atypical in as much as it has a very substantial role as an international examining body in addition to its domestic function as a GCE board.

The University of Cambridge Local Examinations Syndicate (UCLES)

UCLES employs 23 professional academic staff, 31 data-processing staff, and 172 administrative, clerical, and manual staff. In addition, 4,880 examiners are employed on a part-time basis for creating syllabuses; setting, appraising, and editing questions; marking scripts; and fixing standards. A formal description of the syndicate and its committees is provided in Appendix 1.

The data-processing department has an IBM 4341 computer and optical mark reading facilities, and increasing use within the office is being made of IBM personal computers. In addition, some staff, chiefly those involved in research and development work, make use of the University of Cambridge Computer Service's extensive central facilities for data-processing tasks which require software and other services.

The GCE examinations are set twice each year, in the autumn and summer. The latter is the main examination. In summer 1984 UCLES examined 141,438 candidates from 1,593 schools and colleges at ordinary level or its equivalent. Between them, these candidates took a total of 392,138 subject examinations. At the advanced level 41,600 candidates from 880 centers took 66,134 subject examinations.

These are almost entirely for British-based students, but the syndicate has a substantial role as an international examining body. This is at present best summarized by considering two types of examination:

(1) It provides examinations of attainment in various academic subjects, often tailored to the needs of particular countries, very like and often equivalent to the GCE ordinary- and advanced-level examinations. In 1983, 185,210 students from forty-two countries were candidates and between them took 770,960 subject examinations. The main overseas examinations are held in the autumn of each year.

(2) It sets examinations in English as a Foreign Language (EFL), which are designed for non-native speakers and intended as achievement tests for those who have been following a course of study in EFL. These are pitched at three levels for those at different stages in learning: the Preliminary Certificate, First Certificate, and the Certificate of Proficiency. During the 1983-84 academic year 106,314 candidates from seventy-one countries were examined. In addition, the Syndicate collaborates with the British Council in providing the English Language Testing Service as a means of assessing the language competence of

non-native speakers wishing to pursue a course of study or training in colleges and other institutions in English-speaking countries.

This involvement in overseas examining is, of course, atypical of other British GCE examining boards. The syndicate is not, however, atypical in its approach to examining. Its aim is to meet the curricula demands of the schools rather than to impose its requirements upon them. The inevitability of backwash effects from examination requirements on the courses schools teach is widely recognized in Britain. If an externally set test exists and is recognized to be an important qualification for pupils, it is natural that teachers will tailor their efforts towards ensuring that their pupils do well in it. If they fail to do so they will certainly be criticized by pupils and their parents and, in all probability, by the local authority which controls the school. The danger lies in the test assessing a limited set of objectives and tempting teachers to restrict their teaching to those assessed, so constricting the curriculum. Our aim is to set examinations closely linked to the courses schools wish to teach. It is for this reason that we consider it important to provide a choice of syllabuses, within boards as well as between them, constructed to match different views within the teaching profession on the way subjects should be taught. Often, too, we find it necessary to provide internal choice options within an examination to allow the teachers the flexibility to approach their subject in the way they feel is best suited to their own pupils' needs.

Examination syllabuses are constructed and approved by committees involving teachers' representatives and, at the advanced level, university interests too. These bodies also criticize the questions which are set. They are consulted before any developments in their subject area. The examiners who set, moderate, and mark examinations are, in the main, practicing teachers who perform their examining duties as an additional task.

Because it is our aim to make the examinations as valid an assessment of the course of study as possible, our examinations cannot have a set pattern. They may contain whatever tasks we consider necessary to match the curriculum. It is normal to have several components in the examination in a given subject, sometimes covering different subject content or objectives and frequently deploying different assessment techniques. Essay papers, structured questions, practical examinations, coursework, teacher assessments, projects, and multiple-choice questions may all play a part if required.

UCLES GCE Advanced-Level Technology

A detailed description of the aims and syllabus of the course assessed by the advanced level examination in technology is provided in the syllabus leaflet issued to teachers, an extract of which is given as Appendix 2. Its emphasis is upon the capacity to apply technological expertise, and the assessment pattern is designed to reflect this. You will see from Appendix 2 that the examination has four components. Pupils will have followed a course of study in a common core of the subject and will also have studied two from a choice of four more detailed modules, chosen by their schools to match local needs and available expertise and equipment.

The pupils' work in the common core area is central to all that they do. It is assessed explicitly by asking them to produce a folder of approximately six thousand words. In 1984 the topic upon which candidates were asked to work was "the use of steel in the building industry." This required candidates to research the topic and display sound knowledge of it and its relationship to broader issues of technology. The issues they should have considered include: technology and its effect upon man and the environment, energy materials, mechanisms, systems, design, historical developments, safety, financial aspects, and any other relevant broad issues. This component carries 20 percent of the total mark.

The candidates' detailed theoretical knowledge of their chosen specialist modules is assessed explicitly in a three-hour written examination. Part of the 1984 paper is attached as Appendix 3. As you can see, in this particular paper the examiners have chosen to adopt a loosely structured, open-ended question format which gives fairly clear guidance as to what is expected from the candidates, while allowing them considerable scope in formulating their answers. This is essential to convey sufficient detail to make the tasks reflect applied technology, as the course's ethos demands, and to allow for the variation in response that is essential where there can be no single correct solution to the problem set. In other circumstances other assessment formats would be used as appropriate. These tasks are complex ones, and to help ensure that candidates can display their capacity to apply their technological knowledge, the decision has been made in this instance to allow a choice of two questions from four within each section. While question choice is not always justifiable, it is far from uncommon in British examinations. Here, for instance, it is argued that the desire to discover how well students can apply the knowledge they have is more important than the need for blanket coverage of the course's content during assessment. Thus, validity is enhanced at some potential cost to reliability. This component carries 30 percent of the total mark.

The two remaining components are more overtly concerned with students' capacity to apply their skills and knowledge. In a further three-hour written examination testing design skills, candidates must provide a design solution to one problem from a choice of problems. The 1984 paper is shown in Appendix 4. In providing their solution students are expected to display sound knowledge of the syllabus core and the modules that they have studied. They are to analyze the problem and consider alternatives, one of which must be shown in detailed drawings as their chosen solution. This part accounts for 20 percent of the total mark.

Finally, during their course all candidates must undertake a major project dealing with a problem chosen by the student. The report on the project, prepared by the candidate, must include all design work, research, drawings, constructional notes, and evaluation. Students must construct their project. Normally a significant amount of good, standard constructional work is expected to be involved. The project topics must be approved by the syndicate early in the course. This carries 30 percent of the mark.

The examination package is intended to provide due weight to the different parts of the course, while encouraging teachers to require their pupils to learn to apply technology to solve practical problems. The intention is to ask pupils to do things with technology and not just to learn about it. If the examination did not support their approach to the subject and instead concentrated on assessing theoretical knowledge, which is easier to assess tidily and reliably with tightly structured essay or multiple-choice questions, there would be great danger of curricular drift which emphasized the acquisition of theoretical information at the expense of this application.

Marking the Examination

While this may not be the place for a detailed account of the UCLES system for processing examinations, a brief description of what happens once the candidates have taken the written part of the examinations may be of interest.

A chief examiner is responsible for each paper, and in addition to marking some scripts, he or she checks the work of other senior examiners known as team leaders. They, in turn, are responsible for their own marking and for oversight of a team of about five assistant examiners. After the examination the candidates' scripts are posted directly to the examiner who will mark them. The chief examiner selects from his batch some to be photocopied and distributed to all examiners, who then mark these, using a draft mark scheme provided by the chief examiner. These marked photocopies are then brought by all examiners to a coordination meeting where problems are discussed, definitive marks agreed on for the photocopied scripts, and amendments and amplifications to the mark scheme finalized. The team leader (or the chief examiner in the case of team leaders themselves, sends his to a senior team leader and discusses them with him or her. These scripts are returned as quickly as possible, with comments, to the originating examiner. Further scripts can be requested if the team leader is concerned. Otherwise examiners then mark a more substantial batch of scripts and send these to their team leader who provides feedback as necessary. In problematic cases, or routinely in papers where the marking is particularly subjective, this checking by team leaders can continue throughout the marking process. Normally the team leaders' sampling of scripts marked later in the process is relatively light, once they are confident that an assistant is marking consistently. Where examiners are consistently marking high or low they may not be told to alter their behavior; instead their marking will subsequently be scaled to the chief examiner's specifications, with the agreement of the team leaders. Statistical information is provided towards the end of the process to guide such scaling decisions.

Standards fixing is based substantially upon the professional judgment of the senior examiners who will be required, through the chief examiner, to suggest cutoff points for the pass/fail line and the minimum mark for the award of the highest grade. Where an examination and the nature of the entry to it has changed little since the previous year, the percentages reaching the recommended marks will be compared with the results of previous years as supplementary information. Although examiners can and do insist that their judgments should outweigh such normative guidelines where they feel that the general standard of work has drifted over time. Other grade borderlines between these key points are then determined statistically. The dependence on judgmental determination of these key grading decisions is an essential element in the process, and upon it rests the validity of our claim that the examinations are essentially criterion referenced rather than norm referenced. Admittedly much is undefined and much rests upon the experience of the senior examiners and the continuity that this provides, but it enables the system to cope with the setting of complex tasks to the candidates and to react to the need to change the curriculum as a subject develops, while still attempting to maintain constant standards. Following the examination various statistical checks upon the operation of the system are possible. For instance, subject-pairs analysis is used to compare the grades awarded to candidates in different subjects (Forrest and Smith, 1972), although we have learned that the intervention of motivational and other factors should make us treat such information with caution (Newbould, 1982). We have considered various alternative approaches but have not felt them to be appropriate, often because we do not believe that the assumptions which underpin the statistical methodology used elsewhere are met in our context. Thus, the use of a common element in different examinations --in essence, the anchor-test approach-- was investigated by Newbould and Massey (1979) and, perhaps regrettably, found wanting. Such devices, however, can be used to provide information as part of the judgmental process over time. Based as they are on the curriculum, our examinations contain too much diversity to fit comfortably within statistical models. We are loath to tailor our examinations to the statistical tools available, believing this to be allowing the tail to wag the dog.

Following these grading decisions comes the awarding process, during which senior examiners check the work of any doubtful examiners and also reexamine the work of candidates: (1) who are close to key borderlines, (2) who have apparently performed unevenly between papers, and (3) who have notified the syndicate of special difficulties, such as illness or bereavement during the examination period. The award is considered an essential element in our examining process and contributes greatly to enhancing the examinations; it provides reliability of grading beyond that obtained within the initial marking process. Even after the results have been issued, the process of seeking to enhance the fairness of the examination continues. Candidates can request checks for clerical and marking accuracy, and schools can ask for a report on the work of their pupils by the chief examiner -- on payment of a fee. Errors are discovered and corrected sometimes even at this late stage.

Contrast with Other Systems

As can be seen, the examinations used as the basis for university selection in Britain differ substantially from the sole use of objective tests. Our model rejects the notion of the aptitude test, preferring to assess students' performance thus far. We believe that the fact that this performance will necessarily be affected by non-cognitive factors --such as personality, motivation, and application to study-- is a virtue. Such factors will continue to influence the students' performance within higher education, and we do not therefore seek to discount their influence. We fear too the influence of the general achievement test assessing some core of material supposedly followed by all. Where courses differ they are bound to overlap differently with any such core, leading to differential relevance and bias (Newbould and Massey, 1979). Worse, they risk perverting the curriculum by encouraging teachers to focus too closely upon the core because of its role in selection. On similar grounds we deplore any constraints upon the test form, such as a requirement for multiple-choice items because of the ease with which they may be machine marked. We can see no justification for making a test any more artificial than necessary. In these respects we suspect that our practices are in accord with those examinations at present employed in China.

One does of course pay penalties for exercising these preferences, but they are not always as severe as some critics assert, and we are willing to pay the price. Thus, while our types of examination do not easily lend themselves to statistical manipulation for national performance monitoring, we would be unhappy to place much weight upon results from generalized achievement or aptitude tests because of the criticisms of these already made. In a state with less curricular diversity than Britain, it may moreover be more feasible to use open-ended examinations as a means of looking at national performance levels.

In any event, the notion of monitoring national standards may have little relevance when considering examinations for university selection. In the mass higher-education context, variations in test scores may be of interest as indicators of the quality of educational output, but if one is considering a more selective university system, as in Britain, where fewer than 5 percent attend university, they are far less relevant. Again, while our more varied tests are less easily reviewed for the purpose of improving them than are multiple-choice tests, analytic techniques for this purpose have been developed (Nuttall and Willmott, 1972).

In a similar vein we in Britain have not shared the concern felt in some countries over problems of culture fairness in tests. The difference in perspective between testing for aptitude and assessment of achievements may in part explain this, as might differences between legal systems. We have also had the advantage of a reasonably homogeneous racial, cultural and geographic background. We have not sought to ensure that test items are so constructed that different groups of students perform equally well on them and suspect that to do so hides problems rather than solves them. It may be desirable for various reasons to be aware of inter-group differences. For instance, in Britain we have paid considerable attention in recent years to differences in success levels between the sexes in various circumstances, for instance in answering different types of mathematics questions (Wood, 1976), often with a view to taking some remedial action. In some circumstances it might be desirable to discriminate positively. Thus, some British universities might consider relaxing the severity of their standard entry requirements for students whose preparation for GCE examinations they believe has been hampered, say, by attendance at a central city school with few other high-achieving pupils and restricted facilities for advanced work. However, it is not self-evident that one is helped in deciding how best to do this by editing tests to eliminate items giving inter-group differences. To do so necessarily distances the test from its role as an assessment of a given learning program by the introduction of item selection criteria unrelated to content validity.

GCE and University Selection

What use, then, do British universities make of the GCE examination in selecting their students? How does the selection process work? It must be understood that the universities, although government-financed in the main, are independent bodies. Students, wherever they may be resident, may apply for entry to any university and will be considered on their merits. In fact, selection is almost universally in the hands of the teaching department within the university for whose course the application is made. Because students usually wish to apply to several universities in order to maximize their opportunity for success, there is a coordinating mechanism designed to handle multiple applications called the Universities Central Council on Admissions (UCCA, 1984). Within a university department it is common for one member of the staff to take on the brunt of the admissions work, although he or she may be helped by other staff, especially when conducting interviews with prospective students.

The written applications will detail the examinations that prospective students are due to take as well as the results of any they may already have taken. The form also asks for other information. In particular it asks the applicant to describe his or her wider interest and invites her or him to relate significant achievements outside the scholastic world. It includes an assessment of the applicant by an authoritative figure who knows him or her well. Normally this will be the applicant's head teacher, who will coordinate the views of all those who teach him or her at school. Such references are concerned not only with academic potential; they range over character, personal attributes, effort, and willingness to contribute to school and community life. On the basis of this, a department may decide to reject an application or it may decide to invite the applicant to an interview to enable the university staff to form its own impression of him or her. Decisions are made on the basis of a well-rounded judgment, including all these factors in an attempt to decide if a given applicant has both the ability and the character to succeed in the course concerned. Frequently these set a target for the students in terms of examination results which they must achieve in order to gain entry.

Thus, achievement in the GCE is far from the only consideration in the minds of university selectors. They must be sure that entrants are capable of exploiting any talents they have. This is particularly important in higher education systems which, as in Britain, are highly selective on entry and must achieve relatively low failure or drop-out rates. Some evidence points to consistent relationships across different areas of study between personality characteristics and achievement; different types of personality may be most likely to be successful in different fields (Entwistle and Percy, 1973). Well-established links between motivation and study methods and success also exist. (Entwistle and Entwistle, 1970).

It is thus only right that universities should be very concerned to assess such factors, and that examination or test scores alone are insufficient to guide decisions on university selection. We believe, however, that a system like ours, in which motivational factors are an important element in examinations used for selection purposes, can play a useful and important part in ensuring that those who are selected are those best able to profit from the opportunity.

THE SYNDICATE AND ITS COMMITTEES

The Local Examinations Syndicate

The Syndicate is a Committee of the University of Cambridge, an independent charity designed to promote the cause of education. It consists of the Vice-Chancellor (or his deputy) as Chairman, the Treasurer (or his deputy) and eight other resident members of the University, and is assisted in its work by two Councils, the Council for Home Examinations and the Council for Overseas Examinations. The Syndicate retains the right of ultimate control.

The Council for Home Examinations

The Council is responsible to the Syndicate for the management and conduct of its School examinations in Great Britain. It consists of a Chairman (the Chairman of the Syndicate or a deputy appointed by him), four Syndics and not more than twenty other persons. These include three Chief Education Officers, eleven teachers, and representatives of other interests.

The Council for Overseas Examinations

The Council is responsible to the Syndicate and advises them on all matters concerning School examinations overseas and the Examinations in English as a Foreign Language. It undertakes the duties of the former International School Examinations Council and assumes other responsibilities. The Council consists of a Chairman (the Chairman of the Syndicate or a deputy appointed by him), six Syndics and not more than twenty other persons. These include representatives of the Ministries of Education, or of the High Commissions in the United Kingdom, of countries in which the examinations are held, as well as representatives of the British Council, of examining boards overseas with which the Syndicate has close connections, and of other bodies. Relations with individual countries are maintained by visits of members of the Syndicate and its Staff, and by visits of local representatives to Cambridge. The Syndicate is also advised by Subject Panels or Area Committees in many parts of the world.

The overseas School examinations include, in addition to those detailed in these Regulations, special examinations conducted at the request of, and in co-operation with, Ministries of Education or other authorities in various countries.

Subject Committees

There are twenty Subject Committees whose members include, in addition to members of the Syndicate and the Council for Home Examinations, a number of persons appointed by the Council as representatives of the examiners, and of school and University teachers of the subjects concerned. About two-thirds of the members are school teachers. The committees advise the Syndicate and the Council on all matters concerning the subjects under their control. They are responsible for drawing up and revising syllabuses, receiving criticisms and suggestions, and issuing instructions to the examiners.

TECHNOLOGY

9354

GCE ADVANCED LEVEL

Available only in Great Britain and for June examinations

INTRODUCTION

This syllabus has been constructed as a development of the corresponding O level syllabus although a study of this latter is not an essential prerequisite. The syllabus has been so designed to offer students opportunities both for a broad involvement in technology and also to concentrate on detailed areas. The acquisition of a good level of mathematical, scientific, practical and communication skills is necessary in order to achieve the depth of understanding required. The course has been designed for use as a qualification both vocational and for higher education and to give an understanding of the potential and limitations of technology. Considerable emphasis will be placed on the application of scientific principles and technological concepts to solve realistic problems. Whilst there is, therefore, an emphasis on technological capability, the general educational value of technological awareness, analysis, synthesis and evaluation should not be overlooked.

Course Structure

Design is an all pervasive influence throughout every aspect of the course and it is intended that study should be technologically, rather than scientifically, oriented. Candidates are required to study a common core to give them general insight into technology and its effect upon our lives and, in addition, two out of four modules to be treated in depth. The four modules out of which two are chosen for detailed study are described below. Design is given emphasis in the course because it both unifies the common core and the modules and also provides an opportunity for the application of knowledge in the solving of problems.

Module I —Structures
Module II —Automation
Module III —Electronics (Instrumentation)
Module IV —Materials Processing

Additional information about the modules and the design component of the course appear later but it should be stressed that, although there is some natural affinity between Modules I and IV and likewise between Modules II and III, it is thought inappropriate to specify any particular coupling of modules. It is hoped that the free choice of modules will provide flexibility in meeting the needs and interests of schools and students.

A proportion of the final marks will be awarded for a project and time must be allowed for this aspect of the course.

Examination Structure

Paper I (3 hours) Modules, carrying 30% of the total marks

A number of questions will be set demanding a detailed knowledge of the modules. Candidates will be expected to answer four questions, two from each of the modules chosen for study.

Paper 2 (3 hours) Design, carrying 20% of the total marks
Several questions will be set and candidates will answer one. Good answers will display a sound knowledge of the modules, the ability to analyse a problem, consider and present alternative solutions in graphical form and present a final solution in detailed drawings.

Paper 3 Common Core Study folder, carrying 20% of the total marks
Candidates will submit, to the Syndicate, a personal study which should display an understanding of the broad issues surrounding modern technology.

N.B. The topic for the 1984 examination is *The use of steel in the building industry.*

Paper 4 Project, carrying 30% of the total marks
A major project will be undertaken dealing with a problem chosen by the candidate.

Aims of the course

The aims listed below are not intended to be exhaustive nor are they given in any particular order of priority. They are given in the hope that they will stimulate a sympathetic response to, and give some guidance in, the interpretation of the nature of the course.

1. To give an understanding of the design process, its inherent decision making and its application in the solving of technological problems, culminating in self-critical evaluation of the solution against the original specification.
2. To challenge those sixth form students who have the aptitude and ability to become engineers with a course combining academic rigour and technological creativity.
3. To give an opportunity for sixth form students to obtain a technological dimension to their education that will assist them in becoming informed decision makers in a technological age.
4. To exploit inherent creative and inventive talents, by providing a stimulating course that will produce a high degree of technological capability.
5. To provide a course upon which faculties in higher and further education can build.
6. To give sixth form students a body of knowledge and the confidence that will enable them to overcome technological problems by means of workable and workmanlike solutions.
7. To give sixth form students the comprehension and communication skills, both oral and graphical, that will enable them to discuss technological issues with informed and less informed members of the public.
8. To give form students an awareness of the resources and restraints of technology.
9. To give sixth form students an understanding that technology is concerned with working with people and for people.
10. To illuminate the importance of, and provide opportunities for, the application of mathematical and scientific principles.

SYLLABUS DETAILS

A. *Modules*

Four modules are presented from which candidates (or centres) must choose and study two. The four modules available are I, Structures; II, Automation; III, Electronics (Instrumentation); IV, Materials Processing. No restriction is placed on choice.

The modules are important both as a detailed study and as a source of inspiration and information upon which projects may be based. Important theoretical concepts are introduced but it is intended that they should be taught with strong emphasis upon their application in practical situations.

The detailed knowledge gained whilst studying these modules should provide a sound foundation both for later work and in assisting design decision making. Because this interaction is important, it is expected that a good knowledge of the modules will be displayed, where appropriate, in Paper 2, Paper 3 and also in the project.

B. *Design*

Design is the topic which is likely to appear in almost every aspect of the course and the assessment procedure. It is intended that students should be encouraged to consider design in the broad sense (the way in which designers have influenced our world) and in the detailed sense (making decisions in a particular design situation). It is recommended that students be taught design methods and processes and be made aware of their stages, feedback paths and limitations. There are many flow-charts that might illustrate the process but three possible illustrations are given later. Candidates will be expected to demonstrate a sound knowledge of their chosen modules when presenting design solutions and decisions.

C. *Common Core*

The core is intended to serve as an introduction to the course by developing a broad understanding of the technological world in which we live. In particular, candidates will be expected to have achieved a good standard of numeracy and literacy, a sound knowledge of workshop skills and a good knowledge of graphic communication: candidates should also be able to develop and run simple programs in BASIC. Candidates are required to make a detailed study of a topic set by the Syndicate.

Personal study folders dealing with this topic will be compiled by each candidate and sent to the Syndicate for assessment in the year of the examination. The folders should display evidence of good personal research and a sound knowledge of the topic, especially with regard to the broader issues of technology and how they relate to the topic. Issues considered by candidates should include: Technology and its effects upon man and the environment, Energy, Materials, Mechanisms, Systems, Design, Historical developments, Safety, Financial aspects and any other broad issues which may be relevant to the set topic.

The topic for the 1984 examination is *The use of steel in the building industry*.

D. *Project*

A project will be carried out, the choice of topic being made by the candidate. The report on the project, prepared by the candidate, must include all design work, research, drawings, constructional notes and evaluation: this material will be assessed as part of the project.

It is expected that the chosen project will normally involve a significant amount of constructional work and that a good standard of constructional skills will be displayed. Investigational projects with relatively little constructional work may be presented but, in such cases, a written report of very high standard will be expected.

Early notification of the chosen topic should be submitted but acceptance of the project may be assumed unless otherwise notified by the Syndicate.

Details of the modules

Each module contains references to basic laws and/or formulae. They are given as specific areas of knowledge that need to be understood to enable the candidates to apply scientific concepts in technological situations. Emphasis should be placed upon application of knowledge as illustrated by examples of real life technological problems.

Throughout each module, the phrases describing the objectives of the course should be read as a continuation of "Candidates should be able to . . ."

I—*STRUCTURES*

Topic	*Objectives*
1. *MATERIALS*	
Common	State the merits and limitations of various materials and composite materials commonly used in the design of structures, e.g. wood, metal, concrete, reinforced concrete, stone, brick, plastics, laminates and honeycombs.
Appropriate use of materials	Describe factors involved in the choice of appropriate materials for specific purposes, e.g. economic, environmental, strength/weight ratio, corrosion resistance, thermal conductivity.
Properties of materials	Explain what is meant by the terms hardness, toughness, brittleness, ductility, elasticity, ultimate tensile strength and compressive strength. Describe the effects of temperature change on the properties of materials, e.g. creep, strength, ductility.
2. *FORCE*	
Newton's laws Internal and external forces	Explain (i) Newton's First Law and Third Law, (ii) the principles of internal and external forces in a member under load. (iii) the meaning of the terms tension, compression, bending, shear and torsion.
3. *STRESS & STRAIN*	
The Young modulus, modulus of rigidity	Explain the following concepts and use them in the solution of problems related to structural design: direct stress and strain, shear stress and strain, Poisson's ratio, the Young modulus, modulus of rigidity, Hooke's law.

Stress/strain graphs	Draw diagrams showing typical stress/strain graphs for mild steel and duralumin and make design decisions based on the information contained in these graphs.
Factor of safety	Explain what is meant by 'factor of safety' and appreciate its importance in structural design.

4. *SHEAR FORCE AND BENDING MOMENT*

Shear force and bending moment diagrams	Draw shear force and bending moment diagrams for cantilevers and simply supported beams with concentrated loads and uniformly distributed loads. Derive equations for the shear force and bending moment at any section on a beam or cantilever. Select suitable sizes of timber or metal for particular loads, or discriminate between alternative design solutions for the same loading specification.

5. *CENTROIDS AND SECOND MOMENT OF AREA*

Determine the position of the centroid or area of simple sections based on rectangles.

Explain the term second moment of area and perform calculations to determine the second moment of area of common sections, such as rectangular, hollow rectangular, circular, channel, T and I sections.

6. BEAMS AND BENDING

For a beam of uniform cross-section which is subjected to bending stresses:

Neutral axis, moment of resistance	(i) explain the concepts neutral axis and moment of resistance,
Stress variation, maximum stress	(ii) sketch the stress variation across the beam and recognise the position of maximum stress,
Standard bending formula	(iii) apply the standard bending formula $\frac{M}{I} = \frac{\sigma}{y} = \frac{E}{R}$ in the solution of problems involving the design of beams and cantilevers,
Maximum deflections of beams and cantilevers	(iv) given the standard formulae for calculating the maximum deflection of beams and cantilevers, apply them to design situations.

7. *GRAPHICAL STATICS*

Frameworks	Apply the principles $\Sigma F_H = 0$, $\Sigma F_V = 0$ and $\Sigma M = 0$ in the resolution and composition of forces. Use graphical and analytical methods to determine forces in frameworks consisting of many members and carrying loads at two or more nodal points.
Bow's notation	Recognise the limitations of model techniques such as Bow's notation and compensate for these in design situations.
Method of sections	Use the method of sections to determine the forces in particular members of a framework.
Space frames	Use graphical methods to determine the forces in a simple space frame carrying a single vertical load. Select suitable sizes of timber or metal to withstand the forces calculated to be present in any member.

8. *TYPES OF STRUCTURE*

	Sketch and describe (i) common structural components including frameworks, boxes, beams, columns, arches, girders and bearings, (ii) common structures including bridges, cranes, furniture, buildings and dams.

9. *JOINTS IN STRUCTURES*

Methods of joining	Explain how structural members are joined—by bolting, welding, riveting and the use of adhesives. Compare and contrast the effectiveness of different types of joints for a given set of conditions.
Stress distribution	Describe the nature of stress distribution in common joints.
Force transfer	Design a joint that will allow a force to be effectively transferred from one member to one or more other members. Select and use simple methods of reinforcing in design situations, e.g. gussets, ribs, braces and laminating.

10. *FAILURE OF STRUCTURES*

Potential modes of failure	Demonstrate an awareness of potential modes of structural failure, e.g. points of high stress concentration, distortion due to buckling or twisting, laternal instability due to inadeqaute bracing between members. Explain the meaning of the terms brittle fracture, ductile fracture, fatigue, buckling, concrete hinging.

11. *STRUCTURAL DESIGN*

	Appreciate the various requirements of a structure with particular reference to its stability and rigidity under adverse conditions such as flood loading, wind loading and temperature variations. Analyse a simple structure and by intuitive judgement identifying (i) redundant members and (ii) forces and stresses in the members. Design and build a simple structure to satisfy the requirements of a design brief. Carry out tests on the structure and present a critical appraisal of its performance under test conditions.
Design evaluation	Suggest modifications to the design to improve its performance.

12. *TESTING OF STRUCTURES*

	Describe the general principles of the methods used to carry out non-destructive and destructive tests on scale models, inservice structures and specimens.
Strain gauges	Describe the principles of strain gauges and explain how they may be used for (i) comparative testing of components under stress and (ii) monitoring the performance of a structure in service. Explain what is meant by the terms gauge factor and dummy gauge.

Photoelastic stress	Describe the principles of photoelastic stress equipment and explain how polarized light may be used to analyse the stress concentration in simple models.
Brittle lacquers	Describe the use of brittle lacquers to determine the weakest section of a complex structure.
Tensile, compressive hardness, impact tests	Describe the general principles of the method and machine used to carry out tests on metals and non-metallic materials including tensile, compressive, hardness and impact tests.

II—*AUTOMATION*

Topic	*Objectives*
1. *INTRODUCTION TO AUTOMATION*	
	Describe the historical development of automation. Sketch and describe two historic machines or devices which contain elements of automation or mechanisation. Sketch and describe mechanisation arrangements at a local factory or farm.
Mechanisation, automation	Define the terms automation and mechanisation. List the benefits and disadvantages of automation and mechanisation.
Open loop feedback, closed loop feedback	Explain these terms and describe the differences between them.
Hunting, damping, lag	Explain what is meant by these terms.
Simple automation arrangements	Sketch, describe and explain the action of simple automation arrangements found in the home or school such as thermostatically controlled ovens, water cisterns, shower controllers or washing machines.
2. *INDUSTRIAL PRODUCTION*	
Production	Recognise definitions, of and distinguish between, true and false statements concerning productivity, mass production, division of labour, production schedule, flow process chart.
Division of labour	Explain and give examples of how increased task specialisation can increase productivity. List the benefits and disadvantages which result from the introduction of division of labour production.
Labour intensive, capital intensive, work study	Explain these terms.
Industrial systems	Using the terms and concepts in this syllabus, describe in detail the production arrangements at a factory which has been studied.
Automation or mechanised production	Sketch and describe industrial production arrangements which have been mechanised or automated.

ADV/S
9354/1
TECHNOLOGY
Friday
8 JUNE 1984
3 hours
Afternoon

ADV/S
9354/1

UNIVERSITY OF CAMBRIDGE

LOCAL EXAMINATIONS SYNDICATE

General Certificate of Education

TECHNOLOGY

ADVANCED LEVEL

PAPER 1

(Three hours)

Attempt four *questions,* two *from* each *of your* two *selected modules:*

Structures–Questions 1 to 4;
Automation–Questions 5 to 8;
Electronics–Questions 9 to 12;
Materials processing–Questions 13 to 16.

Answer each question on a separate sheet of paper.

A quide to the intended marks for questions or part questions is given in brackets. [ ]

Your pair of answers for a module should be tied separately from your answers for your other chosen module.

Graph paper is available.

This Question Paper consists of 18 printed pages and 2 blank pages.

STRUCTURES

1 A fairground ride, the whole of which rotates around a central pivot, consists of a number of booms each positioned radially about the central pivot and supporting a gondola. Each boom and hydraulically operated ram is pin jointed.

Figure 1 shows that, with the ride in motion and the boom horizontal, the load of the gondola is found to be 10 KN acting at 30° to the vertical.

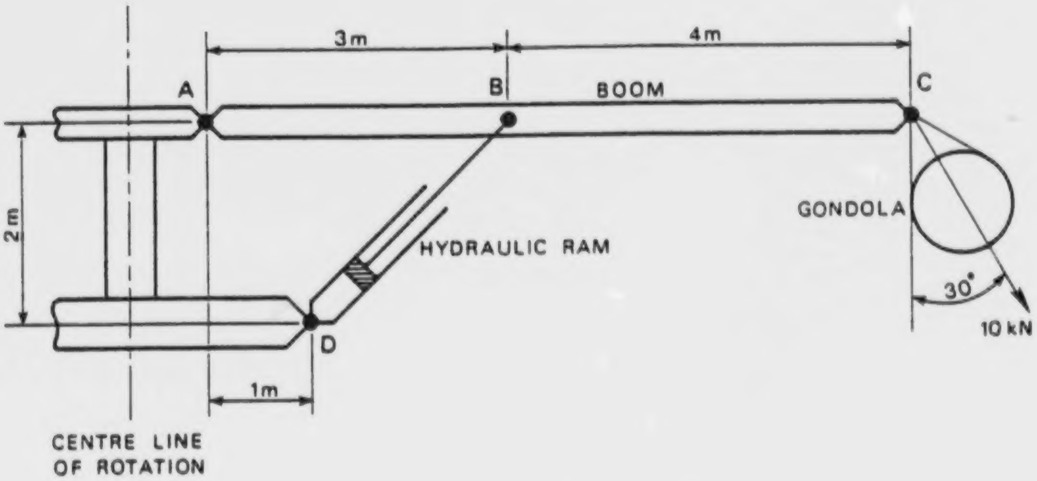


Figure 1

Disregarding any forces originating from the mass of the boom:-

(a) Explain why the load of the gondola does not act vertically downwards, whilst the ride is in motion. [2]

(b) State the nature, direction and magnitude of the reactions at the pivot point 'A' and the piston rod attached at B. [18]

(c) State the position and magnitude of the maximum bending force acting on the boom, whilst the ride is operating in the position shown in figure 1. [3]

(d) If the piston diameter of the ram is 60 mm state the pressure of the hydraulic fluid required to support the boom in the position shown in figure 1. [4]

(e) At what angle would the boom be supported, such that the gondola load, of 10 KN acting at 30° to the vertical, exerted a maximum bending force on the boom. State the magnitude of the bending moment. [3]

2 The framework shown in figure 2 is one of a series set at 2 metre intervals along the external walls of a factory. Each framework is subject to roof and wind loads as shown and also supports two steel conveyor rails suspended below it. Outline details of the conveyor rail are given showing maximum loading, which occurs when the conveyor load of 1 kN passes directly under each framework.

In your answers to the following questions you may;

- (i) assume all joints to be pin jointed;
- (ii) assume the conveyor rail to be simply supported;
- (iii) disregard any forces originating from the mass of the framework and conveyor mechanisms.

(a) Find the nature, direction and magnitude of the reactions occuring in the supporting walls. [18]

(b) The existing wheels of the conveyor are designed to ride on a rail 10 mm wide. If the deflection of the rail, between supports is not to exceed 10 mm, what is the minimum depth of rail, required to support a load of 1 kN?

Maximum deflection $x = \frac{WL^3}{48EI.}$

W = Load.

L = Length of beam between supports.

E = Modulus of Elasticity [E for Steel = 200 GN/m^2]

I = Second moment of Area. [6]

(c) What alterations would you make to the cross-section of a new rail if you were required to improve its design and effective load carrying capacity? Sketch your suggested cross section and show the relative riding position of the conveyor wheels. Give reasons for your design. [6]

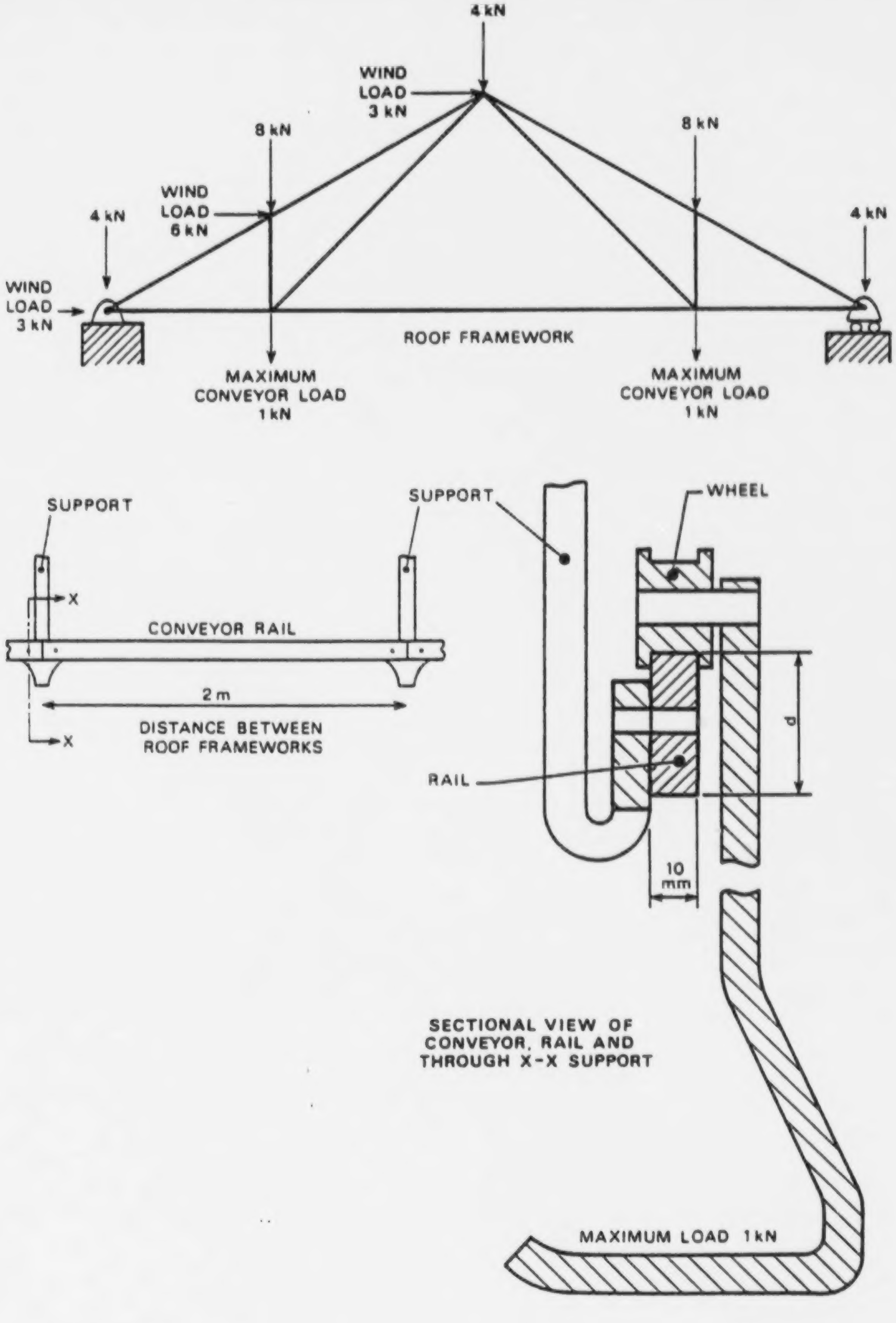

4 kN
WIND LOAD 3 kN
8 kN
8 kN
WIND LOAD 6 kN
4 kN
4 kN
WIND LOAD 3 kN
ROOF FRAMEWORK
MAXIMUM CONVEYOR LOAD 1 kN
MAXIMUM CONVEYOR LOAD 1 kN
SUPPORT
SUPPORT
WHEEL
X
CONVEYOR RAIL
2 m
DISTANCE BETWEEN ROOF FRAMEWORKS
X
d
RAIL
10 mm
SECTIONAL VIEW OF CONVEYOR, RAIL AND THROUGH X-X SUPPORT
MAXIMUM LOAD 1 kN

Figure 2

3 Figure 3 shows a cross-section of the supporting beam of an overhead crane in use in a small engineering factory. The beam, which has a hollow section of uniform thickness, is 4m long and is simply supported at both ends. Allowing 90 N for the weight of the hoist and ignoring the weight of the beam;

(a) (i) calculate, using a suitable safety factor, the maximum allowable load the hoist can be allowed to lift; [13]

(ii) explain your reasons for the safety factor used; [7]

(b) produce design sketches for suitable end fixings to the hollow section beams, which will allow the crane to move, from one end of the factory to the other, on rails fixed to the side walls. [10]

[Yield Stress for Steel = 400 MN/m^2]

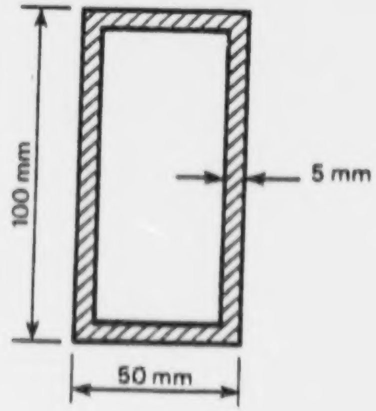


Figure 3

4 After felling, coniferous forest trees are de-barked by a device, illustrated in figure 4, which incorporates a rotating disc and a mechanism which rotates each tree trunk as it is fed past the disc. The tree trunks are held against the rotating disc by spring loaded rollers whilst tungsten carbide tipped teeth, mounted in an offset position on the peripheral edge of the disc, cut or rub the bark off the surface of the trunk. The trunk is rotated whilst being fed past the disc, to ensure that bark is removed from its complete circumference. All protruding branch stumps are sheared off 'flush' whilst passing through the machine.

Experience has shown that discs are failing regularly in two ways:-

(i) The tungsten carbide tipped teeth are 'twisting' and being detached from their mounts.

(ii) Radially orientated cracks develop on the disc.

You are required to;

(a) state where the radial cracks are likely to originate; [3]

(b) suggest reasons for the above failures; [3]

(c) provide sketches and explanations of any design modifications to the disc, and method of feeding the tree trunks if appropriate, which may alleviate the problems described; [11]

(d) describe testing methods that are available to indicate whether your proposed design modifications have been benefical. Recommend one method you think is appropriate for this situation. [13]

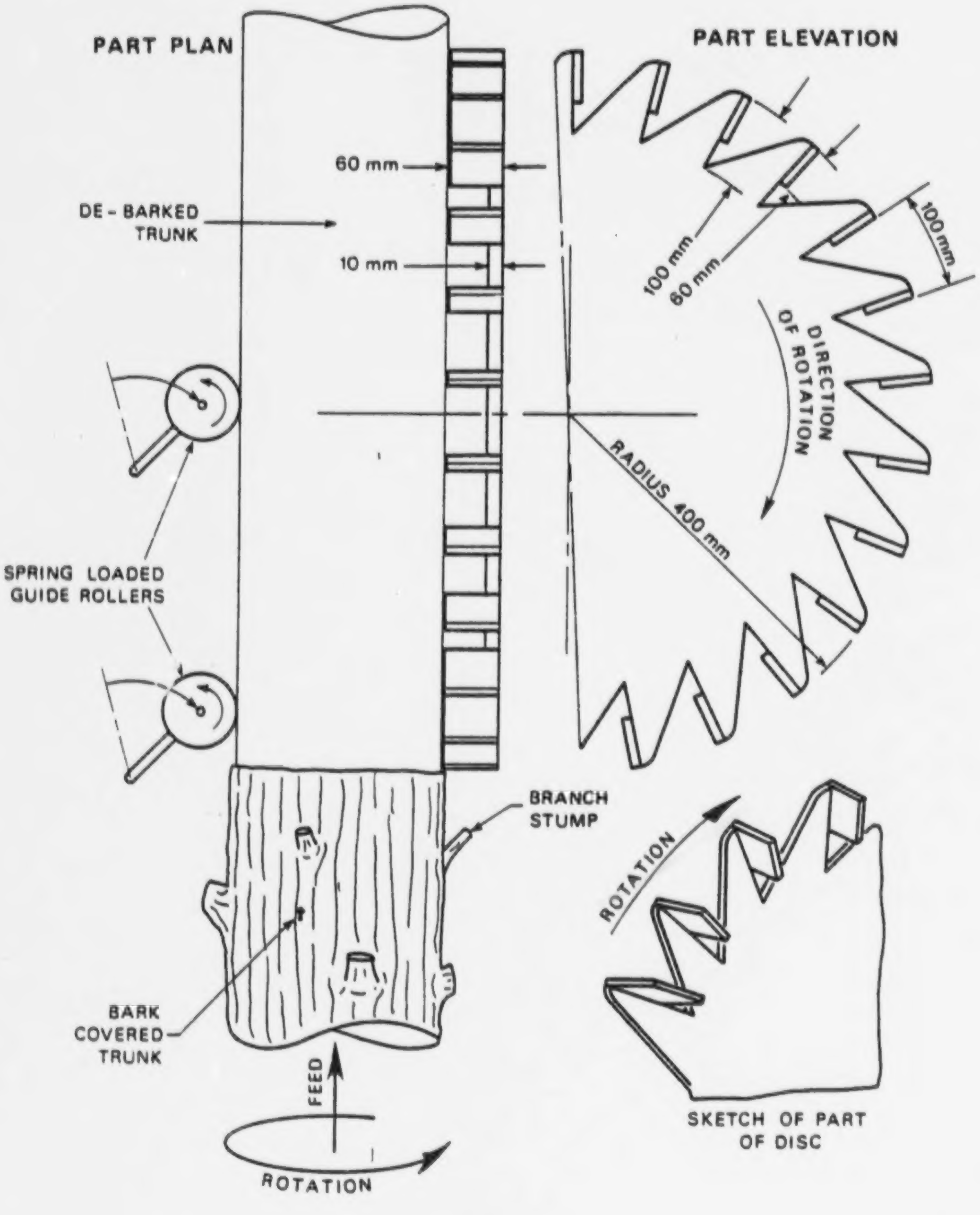

PART PLAN
PART ELEVATION
60 mm
DE-BARKED TRUNK
10 mm
100 mm
60 mm
100 mm
DIRECTION OF ROTATION
RADIUS 400 mm
SPRING LOADED GUIDE ROLLERS
BRANCH STUMP
ROTATION
BARK COVERED TRUNK
FEED
ROTATION
SKETCH OF PART OF DISC

Figure 4

Appendix 4

ADV/S
9354/2

ADV/S
9354/2
TECHNOLOGY
Thursday
14 JUNE 1984
3 hours
Afternoon

UNIVERSITY OF CAMBRIDGE

LOCAL EXAMINATIONS SYNDICATE

General Certificate of Education

TECHNOLOGY

ADVANCED LEVEL

PAPER 2

(Three hours)

Attempt one *question.*

Your answer should contain the following elements:

(a) *a specification of the proposed solution, i.e. a list of criteria that the design must satisfy;*

(b) *brief prescriptions and evaluations of alternative solutions, together with reasons for the selection of the developed solution;*

(c) *relevant explanations and calculations which influenced your decisions;*

(d) *a developed solution, including details of systems of operation, and of components, materials and fixings;*

(e) *relevant working drawings in any of the following forms – sketches, pictorial isometric diagrams, orthographic projection diagrams, circuit diagrams, block diagrams, flow diagrams and graphs.*

This Question Paper consists of 2 printed pages.

1 It is now required for drivers and front seat passengers in most vehicles to wear seat belts.

Design a system and associated hardware that will sense the presence of the person and automatically advance the belt to a position above the shoulder, from where it can be easily gripped and pulled forward.

Include in your design an arrangement that will prevent the engine being started until all front seat occupants are correctly "belted in". If this condition is not met then a warning light must come on on the dash-board.

Your solution must operate automatically upon the first stage turn of the ignition key but must also provide an over-ride facility in the event of an emergency or medical exemption.

2 As part of a manufacturing process a blind hole 100 mm deep and 12 mm in diameter is to be drilled in a brass block 160 mm thick. Design an arrangement to be attached to a suitable vertically mounted drill to enable this process to be carried out automatically and efficiently.

Include in your answer details of your control circuit and feedback arrangements.

Assume that the component is manually clamped and released.

3 Painters face many difficulties of access when repainting upstairs window frames.

Design a modular platform, for use with *sash windows*, that can be quickly and easily assembled by one person from inside the house.

The platform must be large enough to accomodate one painter and allow safe working access to the full width of the window. Access to the platform is to be from the bottom half of the window.

It must also be possible to create a platform link to two or more separate units where different windows are on the same level.

Assume a minimum window opening width of 0.8 metre and a maximum pitch spacing of 1.6 metres.

4 A small village in the Third World has no piped water supply. The nearest water source is a fast flowing stream about 2 kilometres away, the other side of a 50 metre high granite ridge. There is no gas or electricity supply and no road by which petrol or other fuels can be easily supplied.

Design a means by which the stream water can be continuously and automatically supplied direct to the village.

Assume atmospheric pressure will support a column of water 10 metres high.

5 A large covered shopping complex has been constructed of steel columns and girders with glazed infill, to allow maximum daylight penetration to the heated walkways.

Beneath the 14 metre high flat roof, cables for the lighting system, a public address system and a water sprinkler system are housed in ducting.

Design a mobile maintenance platform with a working area of 3 metres by 2 metres, capable of reaching a maximum working height of 12 metres. The platform is to be capable of safe operation by two personnel with minimum disruption to the shoppers. It is to be used for all cleaning, inspection and maintenance work to be carried out above 2 metres within the shopping centre.

References

Bardell, G.S., Forrest, G.M., and Shoesmith, D.J. "Comparability in GCE," Manchester. Joint Matriculation Board. (1978).

Entwistle, N.J., and Entwistle, D.M. "The Relationships Between Personality, Study Methods and Academic Performance." British Journal of Educational Psychology 40 (1970): 132-141.

Entwistle, N.J., and Percy, K. "Critical Thinking or Conformity? An Investigation into the Aims and Outcomes of Higher Education." In Research into Higher Education 1973. Proceedings of the Ninth Annual Conference. London: Society for Research into Higher Education.

Forrest, G.M., and Smith, G.A. "Standards in Subjects at the Ordinary Level of the GCE," Manchester. Joint Matriculation Board. (1972).

Newbould, C.A. "Subject Preferences, Sex Differences and Comparability of Standards." British Education Research Journal, 8, 2, (1982) 141-146.

Newbould, C.A., and Massey, A.J. "Comparability Using a Common Element." Cambridge, TDRU. (1979).

Nuttall, D.L., and Willmott, A.S. "British Examinations: Techniques of Analysis," Slough, National Foundation for Educational Research in England and Wales. (1972).

Wood, R. "Sex Differences in Mathematics Attainment at GCE Ordinary Level," Educational Studies, 2, 2, (1976) 141-160.

UCCA (1984) "This UCCA Business," Universities Central Council on Admissions, Cheltenham.

3

ADMISSION TO HIGHER EDUCATION IN THE UNITED STATES: THE ROLE OF THE EDUCATIONAL TESTING SERVICE

Robert J. Solomon

Introduction

The People's Republic of China and the United States have long shared a belief and a faith in education as the key to building a strong and lasting society. The Chinese, however, have enjoyed a richer and far older educational tradition. In fact, the examinations given to test the educational level achieved by candidates for the Chinese civil service were in common use centuries before there was a United States and have served as models for many nations of the world.

Throughout their history, Americans have viewed education as the key to social progress and individual advancement. At the same time, Americans have always balanced their reverence for learning with a respect for the practical needs of the nation and a very strong spirit of local control and autonomy. As a result, the American system of higher education has evolved in a relatively spontaneous way, free of any systematic central planning or close governmental control.

This irregular pattern was shaped at first by the needs of a new nation, the influence of European culture, and a rapidly changing social and economic order. In the last century, as Americans pushed westward to settle new land, they established colleges based on needs that were different from those of their forebears. In 1862, during the administration of President Abraham Lincoln, Congress passed legislation that allocated more than a million acres for the establishment of tax-supported colleges specializing in courses in agriculture and related fields. Later in the century, in cities such as New York, Chicago, and Boston, large, privately supported universities became centers of scholarship and research along the model of the great European universities.

Meanwhile, beginning in the first half of the nineteenth century, the public attitude about education on all levels was slowly changing from one that considered education beyond the most elementary skills important only for a limited few to an attitude that considered education essential for all the citizens of a growing nation. Eventually, a system of public elementary and secondary schools, paid for and controlled not by the federal government but by each state locality, became a reality. Today, about 75 percent of the young people in the United States complete a secondary-school education.

An important development, which began early in this century and took shape after the Second World War, was the movement to extend tax-supported public education to include the first two years of college. This new idea, which evolved into the junior, or community, college, has turned out to be enormously popular. Community colleges now account for about half the total number of first-time students enrolled in post-secondary education in the United States. However, the great majority of these students do not complete the requirements for the baccalaureate degree. Today, about one in five persons in their twenties graduates from college, although twice that number have been involved for a period in some form of post-secondary education.

The Creation of the College Entrance Examination Board and the Educational Testing Service

As a result of the extensive and heterogeneous growth of public education in a rapidly changing nation, requirements for admission to American colleges and universities in the nineteenth century had become exceedingly complex and disparate. Unlike institutions of higher education in Europe, which were willing to recognize standard examinations or documents of graduation from secondary schools, colleges and universities in the United States each had unique entrance requirements. For example, in 1895 Princeton University required no science preparation, but Columbia University specified preparation in physics and chemistry, and Yale University required botany. Moreover, each university administered and graded its own entrance examinations in different ways.

By 1899 this situation had become so troublesome to the colleges and universities as well as to the secondary schools that representatives from both levels of education formed a membership organization. The new organization, called the College Entrance Examination Board, was responsible for developing and maintaining uniform admissions examinations that tested curricula common to all the member colleges and uniform methods of administering those examinations. However, each member institution was free to use the results of the examinations and to admit applicants according to its own standards.

The initial examination developed and administered by the College Board were Achievement Tests composed of essay questions that tested knowledge of specific academic subjects such as chemistry or geometry. Later in the century, however, there began to be a growing reaction among educators against the rigid curricula then prevalent in the secondary schools. By the nineteen thirties, perhaps heightened by the social upheaval of the Great Depression, this reaction became a full-scale movement. Educators began stressing the importance of general, non-specific intellectual abilities, such as problem solving and critical thinking, that would not only be useful in doing college-level work but would also better equip students to succeed as adults.

It was against this background that the College Board introduced the Scholastic Aptitude Test (SAT) in 1926. Unlike the Achievement Tests, the SAT was designed to measure how well students understand what they read, their ability to reason with verbal and mathematical concepts, their vocabulary, and other abilities that are considered important in college-level work. These abilities are developed over a period of time in and out of school and are not necessarily related to knowledge of specific subjects. The new examination was not an essay test but a multiple-choice test, one in which the student selected for each question, or item, an answer from among several possible choices. The answers a student chose on such a test could be scored correct or incorrect and would not depend on a subjective judgment by a reader of a subjective response by the student. Moreover, scores on these tests were found to be more statistically reliable and, for the growing number of students from diverse backgrounds who took the tests, as valid as the essay tests. That is, scores on these multiple-choice tests predicted the future academic success of students as well as or better than the essay tests.

The College Board subsequently discovered that multiple-choice tests could be scored by machine, resulting in a far more efficient processing system.

About fifteen years later, after its success with the SAT, the College Board changed the format of its Achievement Tests from essay to multiple choice.

Shortly after the Second World War, the widespread use of academic tests in higher education had created such a problem of overlapping and duplication that in 1947 three of the major public service organizations in the United States --the College Board, the American Council on Education, and the Carnegie Foundation for the Advancement of Teaching-- established the Educational Testing Service (ETS) to assume responsibility for the testing programs of these three founders. The objective was to create an organization that had the measurement and research capabilities to set a standard for educational testing in the United States and, at the same time, the logistical and administrative capabilities to provide testing programs nationwide to serve the educational needs of the society, particularly in the area of higher education. The founding and growth of ETS was coincident with the enormous expansion of higher education in the United States in the years after World War II.

The Current Scene

Today, eight decades after the establishment of the College Board and almost four decades after the founding of ETS, higher education in the United States has become more complex and diverse. Of the approximately four million students who enter secondary education each year in the United States, about three million, or 75 percent, graduate, and of this group, slightly more than 50 percent go on to a two- or four-year college. Almost all college applicants are admitted somewhere, although not necessarily to the institution of their choice. Of all those admitted to college, about half eventually receive the baccalaureate. About 25 percent of those with a baccalaureate degree eventually enroll in graduate school.

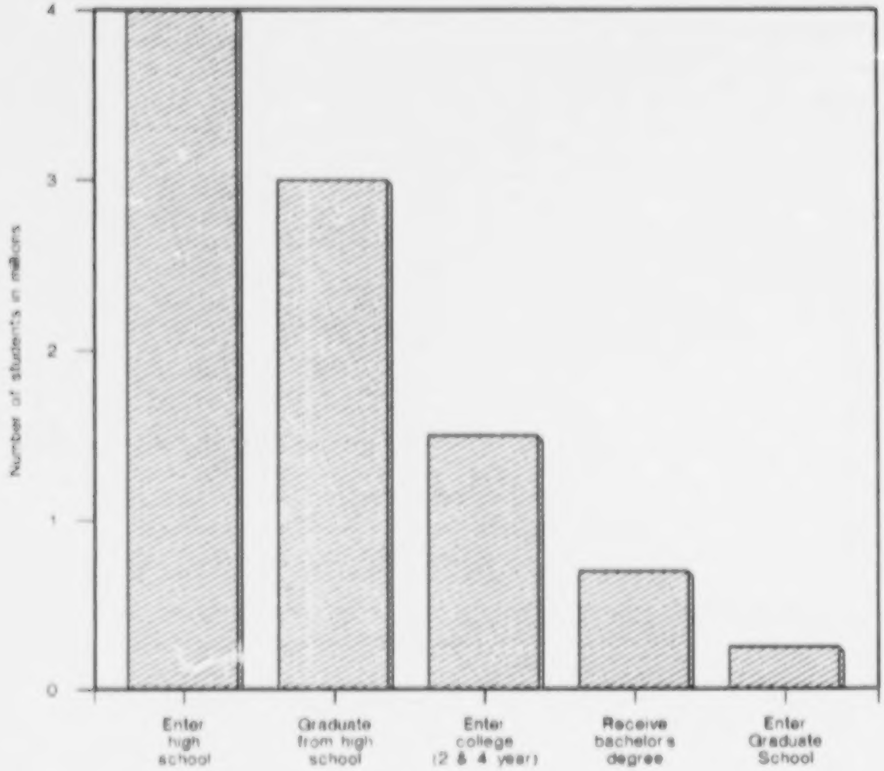
A Wide Diversity of Programs and Institutions

About three thousand post-secondary institutions in the United States serve a total population of about 237 million, of whom about 12 million are enrolled in post-secondary education each year. They reflect the values, diversity, and demands of the society itself. They include large public universities with many undergraduate programs; professional schools of medicine, business and law; graduate schools for scholarship and research in the arts and sciences; small liberal arts colleges; and graduate schools that specialize in curricula to train prospective teachers, scientists, engineers, accountants, and many other professions.

Nearly one thousand two-year colleges offer an enormous variety of curricula, many with a vocational orientation. These include night schools for those who cannot attend during the day and courses of general studies for those who do not which to pursue degrees. Courses are also given through correspondence and television, although the easy access to college facilities has tended to limit the demand for these less traditional approaches.

No one definitive system of admissions guarantees entrance to, or restricts students from, any one college or university in the United States. Rather, what appears on the surface to be an anarchic system of curriculum requirements, grade requirements, and test standards serves to create a distribution system in which a college applicant is encouraged to follow one path and, often, is

Figure 3.1. U.S. Students in Secondary and Post-Secondary Schools



discouraged from following another. Within this system, admissions tests close some doors but open others.

Virtually everyone planning to continue beyond secondary education in the United States takes a test of some kind. In most instances, the test results are used for making admissions decisions, but many colleges, such as those community colleges that admit all applicants, use the results to place students in appropriate courses according the their academic strengths and weaknesses. The best known and the most prestigious of these tests are those of the College Board, but there are other widely used tests, most notably those of the American College Testing Program (ACT), which was established in 1959.

The ACT tests are required mostly by large, publicly supported universities, primarily those in the central southern regions of the United States.

Admissions policies and standards range from open admissions (admitting all who apply) to highly selective admissions (admitting approximately one out of every ten to fifteen applicants). Admissions policies are set in a variety of ways:

- Community (two-year) colleges usually follow a policy of open-door admissions
- At the other extreme, a number of well-known private universities and a few publicly supported ones adopt an institutional policy that restricts admission to the best prepared students in the nation and results in a highly selective admissions policy.
- A few states, such as California, set specific admissions policies and then designate their community colleges, state colleges, and state universities to serve different levels of students (in terms of academic preparation) according to those policies. In California, the state university system is very selective, but the community colleges admit all applicants.

The admissions staff of a college or university is usually composed of a director or dean and a staff who review applications for admission, frequently interview applicants, and visit secondary school students in their schools to encourage then to apply and to inform them of application procedures. There is usually one committee made up of faculty and administrators who develop admissions policies and review their implementation. Often, especially at the highly selective colleges, members of the faculty help to decide on applications for admission.

The choice of a college is initially the student's. Whatever the admissions policy of a particular institution, the student selects the college more often than the college selects the student. Because of this self-selection factor, it sometimes appears that the great majority of students go to the institutions of the first choice. Actually, the student's advance knowledge of what standards are required for admission to various institutions has a great deal of influence on where they apply and, therefore, where they are eventually admitted. The sources of this advance knowledge are the colleges themselves, organizations such as the College Board and ETS, secondary school counselors, and, of course, information and hearsay from friends, schoolmates, and family members.

In summary, the admissions process in the United States is more a system of distribution than a system of selection. There are enough places in post-secondary institutions for every applicant. Where or whether the student goes depends on individual interests and preference, previous education, academic ability as reflected in course grades and test scores, family background, and ability to pay. In the United States, the ability to pay is not inconsequential. Although public and private sources of financial support are available, and although tuition at many publicly supported institutions is relatively low, the annual tuition and fees at the most selective private universities amount to more than half the annual income of a middle-class American family. This is why higher education in the United States has such an extensive system of governmental, private, and institutional student financial aid that includes scholarships, grants-in-aid, work-study programs, and subsidized loans. In addition to government funds which directly support public colleges and universities, 47 percent of American students in higher education receive some form of direct financial aid, and 68 percent of those enrolled in private colleges and universities receive such assistance.

The Admissions Process

The admissions process begins with the student who, often with the advice of parents, teachers, school counselors and/or friends, decides which colleges to apply to. Many students apply to only one or two colleges. The typical American student is likely to go to a college less than 50 miles from home. Many, however, apply to several colleges or universities, some of them far away from home. They send their applications directly to each college and university.

The formal admissions process begins during the student's final year of secondary school, usually in the autumn, a year before starting college. The student requests an application form from each institution he or she is interested in attending and determines from the institution what its requirements are for admission (including taking admissions tests such as those given by the College Board or other testing agencies). Usually the applicant will be required to submit the application to the college by January or February. The student will have arranged to take the required tests before that time so that, along with the completed application, the college will simultaneously receive the test scores from ETS and a record of the student's grades, teachers' appraisals, and other information from the school. The admissions office studies all the information and arrives at a decision in the early spring. However, students and their families who are most committed to a college education often begin planning for college application years before the final year of secondary school.

Of all the information the colleges receive about their applicants, the students' courses and grades in secondary school are given the greatest attention, particularly if the schools involved are known to the colleges. This is because grades have been proven to be a very good predictor of a student's academic performance in college. Clearly, a student's previous education has a direct relationship to his future education.

Scores on admissions tests such as the College Board's SAT and Achievement Tests and those of the American College Testing Program provide further evidence of the student's academic ability. Generally test scores are almost as predictive of a student's future academic success as the student's previous academic record. The combination of previous grades and test scores is more predictive than either one alone.

In addition to grades and test scores, admissions officers consider reports of the student's non-academic activities, both in and out of school, evidence of motivation to succeed in college, and recommendations of teachers and counselors. For most students, however, none of these is as important as grades and test scores.

How the College Board and ETS Serve the System

Since the founding of the College Board in 1899, the number of its member institutions has grown from twelve schools and colleges to more than 2,500 public and independent colleges, universities, secondary schools, and education associations. The purpose of the Board remains basically the same: to provide a forum and services through which the American educational community can facilitate the transition from secondary to higher education.

Today, through the 1,175 colleges and universities that are Board members differ widely in their entrance requirements, nearly all of them require their applicants to take the SAT. Two hundred colleges and universities that are generally more selective than the rest also require achievement tests. In addition to these admissions tests, which are taken by more than 1.5 million students throughout the world each year, the College Board offers a number of other testing programs and services, all designed to make the transition from secondary school to college somewhat easier. The Preliminary Scholastic Aptitude Test, for example, is a guidance instrument designed to give students in the eleventh grade an indication of how they will perform the College Board entrance examinations that they will take in the twelfth grade. Scores on this are used in counseling students about post-secondary education, to help them decide whether they will go to college and to which institution they will apply. The Advanced Placement Examinations are taken by students who are in advanced courses in secondary school to earn credits toward graduation or advanced placement, or both, in the colleges or universities they enter. In contrast to those programs, which serve students who move in traditional ways from school to college, the College Board offers the College-Level Examination Program for adults who have not completed their college education, for students transferring from one college to another, or for people who never went to college at all. Scores on these examinations are used to earn credit at a university or advancement in one's career.

An essential function of the College Board for many years has been to serve as a forum for discussion and debate on matters of college admissions. The annual meeting of the board's member institutions, the several regional conferences held during the year, the colloquiums and workshops, and the Board's publications all serve as a means of communication among representatives of secondary schools and institutions of higher education.

Unlike the College Board, which provides direction and coordination for its member institutions, Educational Testing Service provides technical services in the fields of measurement and research for all levels of education. These include administering the testing programs of its founders as well as those of many other organizations, developing tests, serving as a source of information and instruction for schools, colleges, and individuals on testing and selection, use, and interpretation of test scores, and conducting research in the fields of education and measurement.

Both the College Board and ETS give substantial attention to improving the understanding and use of tests and test scores. Standardized tests are instruments derived from an extensive body of technical knowledge. Not all of these who use tests have the opportunity to acquire that technical knowledge. Nevertheless, the limitations of tests, as well as the helpful information they can provide, need to be well understood by those who make policies involving the use of tests as well as those who receive and use test scores. In the United States, this includes everyone from government officials and university presidents to admissions directors, school counselors, teachers, students, parents and the general public. The College Board and ETS and the other organizations for which ETS develops and administers tests, attempt to increase understanding of tests and improve the use and interpretation of test scores through a variety of channels, including extensive publications, workshops, and conferences. Each person who takes a test or receives a test score is given material that explains the purpose of the test, the specific content, how the test is administered and scored, how to interpret the scores, and the limitations of the scores. In addition, for those who are expert in tests and testing, technical manuals and research reports are made available.

The testing programs administered by ETS span all levels of education from the primary grades through graduate school. In addition to the various testing programs of the College Board, ETS administers tests for admission to secondary and graduate school as well as the National Teacher Examinations, which are widely used for the certification and licensing of teachers. ETS also develops and administers licensing and certification tests for more than fifty occupations and professions.

Like many private colleges and universities in the United States, ETS is a nonprofit organization that belongs to no one person or group of persons. It is chartered under laws that were especially created to foster non-governmental actions for the public good. In the United States, many hospitals and colleges are nonprofit organizations, as are the Carnegie, Ford, Luce, and other foundations.

Although it administers programs for the federal government, ETS is entirely independent of its control. It is governed by many people in many ways. Its policies are directed by a board of seventeen trustees who represent many areas of education as well as other fields. The policies of the trustees are carried out by eighteen officers who are employees of ETS. Numerous advisory committees, made up of educators, testing and guidance leaders, and other experts in various fields, work closely with ETS to help plan its activities.

The Educational Testing Service employs a staff of 2,200 including about 650 professionals, most of whom have advanced degrees in measurement, education, computer science, or management. Its services are accessible to schools, colleges, and other educational institutions through eight Field Service Offices throughout the continental United States and Puerto Rico as well as through its headquarters, in Princeton, New Jersey.

About 90 percent of the income of the organization is derived from fees paid by test takers. In 1983-84, out of a total expense budget of $143 million, ETS spent $119 million (or about 83 percent) on the development and conduct of testing programs and $18 million (or about 13 percent) for research and development.

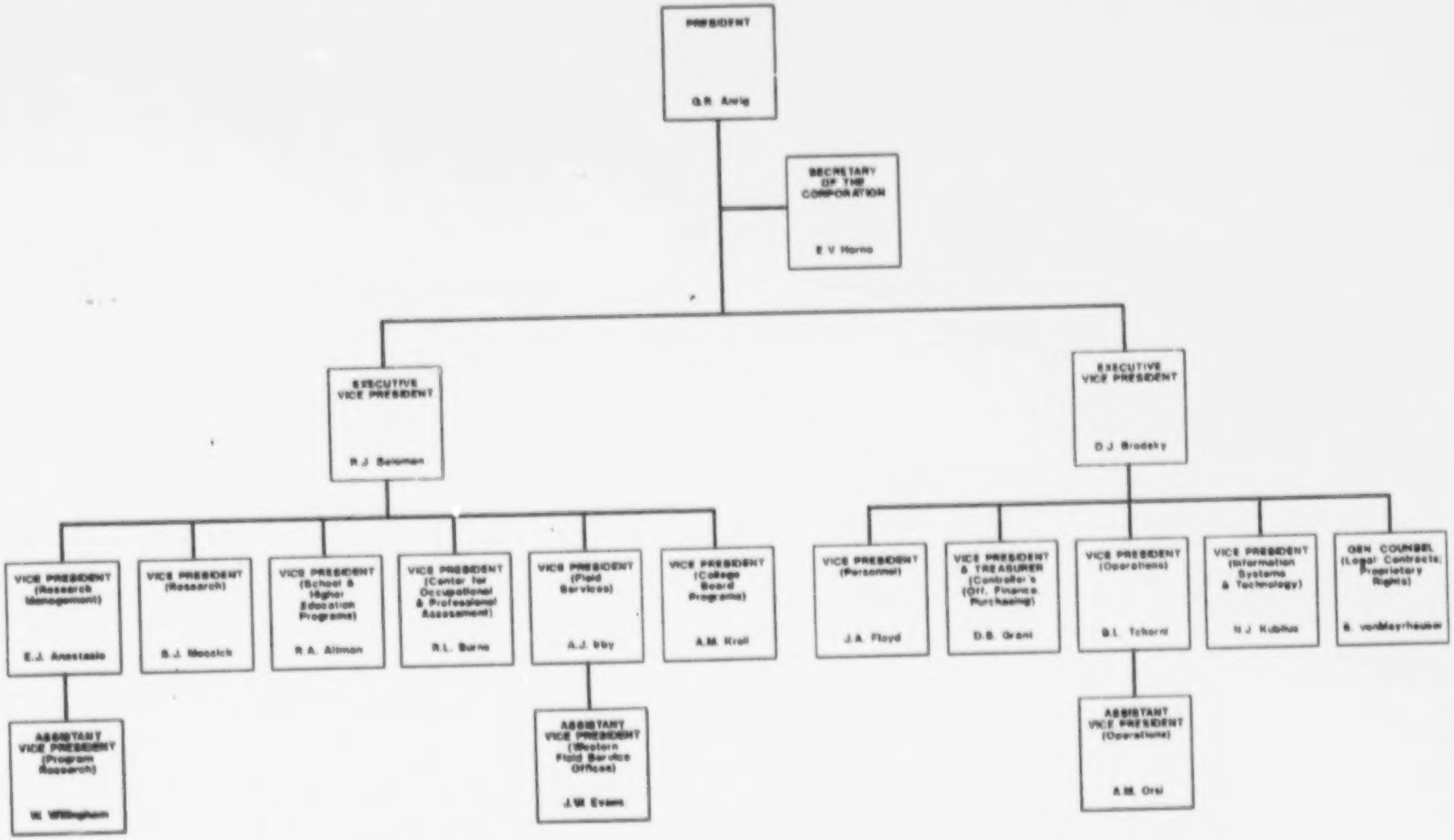
Developing Tests

The ETS test development staff includes nearly one hundred highly trained specialists in academic disciplines as well as psychometrics. These experts, most of whom are former teachers, work with committees of scholars and teachers from schools and colleges across the nation to identify the purposes that each test is expected to serve, specify the content of such test, and determine the kinds of abilities or levels of thinking to be tested. The members of the committees represent a variety of institutions, regions, and points of view as well as the breadth and depth of scholarship required. They not only help to plan tests but also write and review the questions and review the final drafts.

Once a decision to develop a test has been made, test developers ask a number of important questions, such as:

- Who will take the test and for what purpose?
- What can test takers be expected to know about the subject being tested?
- How should test takers be able to use this knowledge?

Figure 3.2. Educational Testing Service (ETS) Corporate Structure


PRESIDENT
G.R. Anrig
SECRETARY OF THE CORPORATION
E.V. Horne
EXECUTIVE VICE PRESIDENT
R.J. Solomon
EXECUTIVE VICE PRESIDENT
D.J. Brodsky
VICE PRESIDENT (Research Management)
E.J. Anastasio
VICE PRESIDENT (Research)
S.J. Messick
VICE PRESIDENT (School & Higher Education Programs)
R.A. Altman
VICE PRESIDENT (Center for Occupational & Professional Assessment)
R.L. Burns
VICE PRESIDENT (Field Services)
A.J. Irby
VICE PRESIDENT (College Board Programs)
A.M. Kroll
VICE PRESIDENT (Personnel)
J.A. Floyd
VICE PRESIDENT & TREASURER (Controller's Off., Finance, Purchasing)
D.B. Grant
VICE PRESIDENT (Operations)
B.L. Tchorni
VICE PRESIDENT (Information Systems & Technology)
N.J. Kubilus
GEN. COUNSEL (Legal Contracts; Proprietary Rights)
E. vonMayrhauser
ASSISTANT VICE PRESIDENT (Program Research)
W. Willingham
ASSISTANT VICE PRESIDENT (Western Field Service Offices)
J.W. Evans
ASSISTANT VICE PRESIDENT (Operations)
A.M. Orsi

- What kind of questions should be in this test and how many of each kind?

- How long should the test be?

- How difficult should the test be?

The answers to these questions form the specifications, or blueprint, from which the test items are written and the final editions, or forms, of the test are made. Actually, the answers to most of these questions involve statistical as well as subject-matter considerations.

Typically, the test committee that draws up specification for a test also assumes the responsibility for writing the test items. For standardized tests --those that are administered under the same, standard conditions at all locations where students go to take the tests-- ETS most frequently uses the multiple-choice question format. Although this type of question, or item, can be processed more efficiently than most kinds, it is also far more difficult to construct than it might appear. The author of the multiple-choice item must word the question and the answer choices in such a way that those who know what is being asked can answer correctly and those who are less knowing will, in effect, be misled by their own ignorance to choose an incorrect response. In this way, students who have not learned the subject or have misconceptions about it will be differentiated from those who are knowledgeable about the subject. In a multiple-choice question the closer the incorrect choices are to the correct answer, the more difficult the question becomes.

Throughout the question-writing process, an ETS test development specialist assists the committee by offering advice on the construction of the questions and by reviewing the questions as a technical editor. As a final step, the committee selects from among all the questions it has written those it considers to be of sufficiently high quality to include in its test.

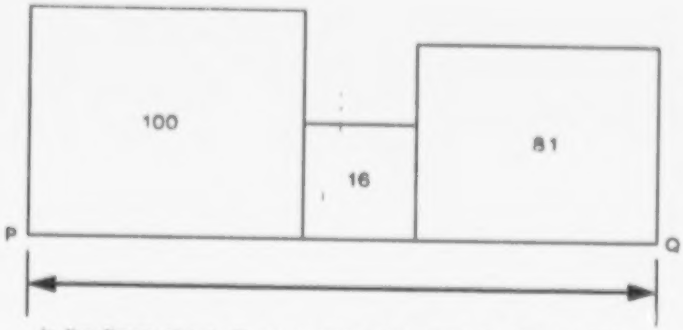
The questions are then tried out experimentally in what is call a "pretest" administered to a sample of students similar to those for whom the test is being devised so that developers can determine (1) the difficulty level of each question; (2) whether individual questions are ambiguous or misleading; (3) whether questions should be revised or discarded; (4) whether incorrect alternative answers should be replaced or revised.

The mathematics item shown below illustrates how this process works. This question is similar to those included in tests of mathematical aptitude for typical secondary school students. An important objective here is to measure reasoning ability that the student can apply to new learning situations throughout college. Therefore, there is a need to avoid using familiar and specific textbook material while at the same time keeping the content level of the question within the context of those subjects the student has already mastered --in this case, geometry and algebra. If this objective is met, then the focus of the aptitude test is where it should be --namely, on ingenuity in solving novel problems rather than on the recall of memorized procedures. The student is given about twenty-five such problems to solve in thirty minutes.

This question involves procedures that the student is unlikely to have encountered in school textbooks; the content knowledge is certainly not beyond what is taught in earlier grades. The solution proceeds as follows:

The lengths of the three sides of the squares are 10, 4, and 9, reading from left to right. Therefore, initially PQ is to be 21; the amount of reduction of the side of the middle square must be 2. The middle square must consequently be reduced from an area of 16 to an area of 4; that is, it must be reduced by 12 square units.

Sample ETS Mathematics Item



In the figure above, three squares with areas of 100, 16, and 81 lie side by side as shown. By how much must the area of the middle square be reduced in order that the total length PQ of the resulting three squares be 21?

(A) $\sqrt{2}$　(B) 2　(C) 4　(D) 8　(E) 12

Table 3.1. provides a statistical analysis of a random sample of three hundred students from the total population who tried this question. The analysis is typical of the procedures ETS follows with all test questions.

Table 3.1. Statistical Analysis of Test Takers

	Students Classified by Total Test Score				
Responses	Lowest Fifth	Next Lowest Fifth	Middle Fifth	Next Highest Fifth	Highest Fifth
Omit	30	12	12	4	1
A	8	9	8	2	1
B	11	18	8	10	3
C	4	3	4	3	4
D	2	3	3	1	-
*E	5	15	25	40	51
Total	60	60	60	60	60

Percentage of total group of 300 students answering correctly	45%
Correlation between success on this question and total score	66%

*Correct answer

Of the fifty secondary-school students in this sample who omitted this question, fifty-four were in the lowest three fifths of ability as measured by their total performance on the test. The fifty who chose answer choice (b) were students who started toward a correct solution but who stopped with the length of the new side of the middle square. The eighteen selecting choice (c) seem to have proceeded as far in their solution as finding the new area of the middle square and, on the average, this was the most able of the groups missing the question. Choices (a) and (d) result from taking wrong directions in the solution of the problem and, as expected, these were selected by the least able groups.

The question was of moderate difficulty for the students taking the test. (In this regard, it is important to keep in mind that in order to optimize the differentiation among students taking the test, the typical average difficulty of the questions in such tests is between 50 and 60 percent correct.) Moreover, the question itself differentiated between able and less able students. The correlation between success on this question and total test performance was 0.66, which is relatively high. Typically, the correlation between individual questions and total test score is about 35, although the correlation depends very much on the homogeneity of the content among the questions and needs to be interpreted in that light. Nevertheless, the correlation is important because on it rests the reliability of the total test score.

Before and after each pretest, the assembled questions are given a sensitivity review to ensure that they reflect the many cultures of American society and that references to minorities and women are positive and appropriate. Moreover, each test is reviewed to ensure that any work, phase, or description that could be regarded as biased, sexist, or racist is removed.

In the final phase of test development, test specialists analyze the data from the pretest results and once more choose from among the questions those that are most appropriate for the subject matter and for testing specific skills. These questions are then assembled into a test following statistical specifications that will ensure that appropriate overall difficulty level and power of the test to differentiate between able and less able students.

After the test is assembled, it is reviewed by other specialists, committee members, and sometimes by outside experts. No test can go to press until the person responsible for it certifies that at least three different people have independently agreed to the correct answers to every question.

After the test has been administered, but before final scoring takes place, a preliminary statistical analysis of each question is carried out based on a sample of several thousand answer sheets. If a problem is found in any of the questions, corrective action is taken before all the answer sheets are scored and the scores are reported. Finally, the scores on each form of a test are, through a statistical process, put on the same common-score scale as previous scores on that test. This means that a score on the edition of the SAT taken in one administration of the test, for example, will mean the same as the score of a different edition of the SAT taken at a different administration. This process of scaling and equating is essential to enable college admissions officers to compare the test scores of different students.

How ETS Administers Tests

In 1983-84 ETS was involved in the administration of tests to about eight million test takers in more than one hundred testing programs throughout the United States and more than one hundred fifty other countries and regions.

Such large-scale operations are possible only because procedures have been developed and refined over the years, and staff members have been trained in their use. Establishing centers where students can take its tests and staffing those centers is a major task, particularly for the larger testing programs such as those for the College Board. Each year, more than ten thousand such centers are established in the United States and other countries involving about thirty-five thousand administrations. To accomplish this means coordinating tests administration activities with more than fifty-five thousand supervisors, associate supervisors, and proctors who are responsible for administering the tests at those centers. As many of you know, through the China International Examinations Coordination Bureau of the Ministry of Education, seven test centers have been established in the People's Republic of China for the administration of ETS tests such as the Test of English as a Foreign Language (TOEFL), the Graduate Record Examinations (GREO), and the Graduate Management Admission Test (GMAT).

Students are informed about ETS testing programs by means of a series of publications. The College Board's Admissions Testing Program (ATP), for example, publishes an annual Student Bulletin containing information about the tests, the times, dates, and places where they will be given, and a form the students fill out to register for the tests. Those students who register receive free booklets to help them prepare for the tests.

Once their registration forms and fees reach ETS, test administration staff assign the students to the test centers the students have chosen or to one nearby if those preferred are not available.

In 1982-83 more than 1.5 million SAT tests (in addition to the other tests given by ETS) were taken at approximately 3,800 test centers, about 500 of which were located in foreign countries and regions on six continents.

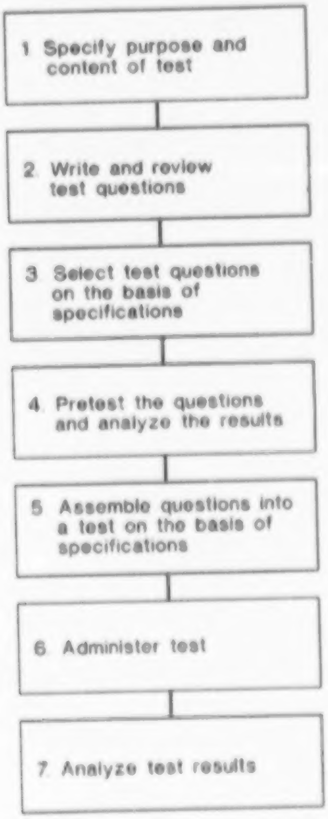
Within four to six weeks after a test administration, the test scores are sent to the students and to the colleges and universities they designated on their registration forms. The task of translating the response marks on the answer sheets into interpretable score reports is done by electronic scoring machines and computers. To ensure accuracy in this scoring and reporting operation, ETS maintains a quality control system that rejects any answer sheets the scoring machine is unable to "read." These answer sheets are then scored by hand.

Electronic Support Services

The period of 1950 to 1970 was one of extraordinary increases in the numbers of people tested by the College Board. It was fortunate that those two decades were also the years in which electronic technology came into its own. Without high-speed data processing, the expansion of services to these much larger student populations would have been inconceivable.

To meet its data processing needs, ETS maintains an entire division of more than three hundred staff members and several computers--two large mainframe computers (an IBM 3083 BX and a National Advanced System 8053) and several smaller ones--which are housed in 17,000 square feet and are in operation twenty-four hours a day. The electronic machines that score the sheets on which students mark their answers to tests are special-purpose computers. By means of a light source, they "read" the pencil marks that students have entered on their answer sheets.

Figure 3.3. Steps in Developing a Test


1 Specify purpose and content of test
2. Write and review test questions
3 Select test questions on the basis of specifications
4. Pretest the questions and analyze the results
5 Assemble questions into a test on the basis of specifications
6. Administer test
7. Analyze test results

A Commitment to Research

As the nation's largest independent, nonprofit testing organization, ETS has, from its beginning, been conducting research not only in testing and other forms of assessment but also in the ways they might be used to improve the teaching-and-learning process.

Today, ETS conducts the most extensive and diversified program of basic and applied measurement research in the world. Research scientists at ETS are studying how technology, such as computers, can be applied to improve testing and instruction. As it has for many years, ETS is very much involved, too, in discovering more about how children learn; in finding new and better ways of testing; and in conducting studies to determine whether its test are appropriate measures for both sexes, and for the different races, ethnic groups, and handicapped students who take them.

An important part of this research effort has to do with the transition of students from secondary to higher education. For example, one recent study, conducted jointly with the College Board, investigated how factors other than students' secondary school records and test scores affect success in college. Research scientists identified the characteristics that are regularly associated with students who are successful in college and proposed ways in which admissions officers can judge such characteristics more effectively.

Although ETS is well known for its research on psychometric theory and statistics, much of its research is also of direct service to educational institutions. For example, ETS administers for the College Board a free Validity Study Service for colleges that wish to evaluate how well their admissions data (test scores and secondary school record) help them predict the academic performance of their students. Such information has proved helpful to admissions officers in deciding how to use test scores along with other academic aspects of the secondary school record to predict a student's chances of academic success. For example, in recent years, when higher grades were easier to earn in secondary schools in the United States, the validity research showed how these grades were less useful (less predictive and the test scores more useful in many admissions situations.

Two services introduced by the College Board in the 1970s were in the Student Descriptive Questionnaire (SDQ) and the Summary Reporting Service. The SDQ contains questions about students' backgrounds, academic records, extracurricular activities, plans for future education, and other information. Summary reports to colleges and schools combine the information on students' test scores with the personal information about their backgrounds taken from SDQ. Thus, for example, colleges can find out how many of their enrolled students who scored at a certain level on the SAT come from certain regions of the United States or hope to go to graduate school, or would like to specialize in the study of science or languages or mathematics during their college career.

On a national level, the SDQ data provide valuable information about trends among the college-going population, changes in patterns of college preparation, of career interests, and any other aspects of the student population that is important for policy making purposes.

Conclusion

Between two and three million students apply for admission to three thousand American colleges and universities each year. This creates a difficult problem for those who must make decisions about who will and will not be admitted. This difficulty is compounded by the diversity of American secondary schools and the quality of their educational programs. It is difficult for admissions officers to predict, for example, how a student will perform who has an excellent record but is from an unknown school. Admissions officers often find it equally difficult to make decisions about a student with a mediocre record from a known school with exceptionally high standards.

It is in this context that admissions tests such as those developed for the College Board by ETS have played a major role. Unlike the academic standards, and methods of grading, scores from these tests, which are given under standard conditions throughout the nation each year, represent a common measure of ability that can be used to compare applicants from schools everywhere. A student's school record is a very accurate predictor of how the student will perform in college. With the addition of test scores, the school record is an even more powerful and accurate means of prediction.

Because test scores are a useful and fair method of discovering, comparing, and selecting students from a large population of applicants, their use has become widespread throughout the United States. As stated earlier, literally everyone who plans on higher education in the nation today takes some kind of test. This has imposed a burden of responsibility on the Educational Testing Service and the other organizations that develop and administer educational tests. We are obligated not only to produce tests that are accurate, fair, and appropriate for their purposes, but also to be certain that the results of our tests are properly interpreted and applied by the institutions that use them.

Despite the progress that has been made in improving and facilitating the selection of students for higher education in America, difficulties remain. As one former admissions officer wrote, "Sorting and guiding individuals in society to maximize their potential and to fulfill society's needs is...much of the business of the educational world." Measurements such as rank in class, high-school grade averages, and test scores can help, but they cannot do the entire job. Ultimately the decisions must be made and the responsibility for them must be borne by conscientious human beings.

4

PUBLIC EXAMINATIONS IN AUSTRALIA

John Philip Keeves

An Overview of Australian Secondary-School Examinations Systems

Most Australian states are currently making significant changes in their conduct of public examinations at the terminal secondary-school level. However, the scene across Australia is relatively complex because each state system is approaching its problems from a different perspective and with different traditions. In order to understand the problems and the widespread debate, it is necessary to recognize that in Australia education is primarily the responsibility of the seven state governments, including the Northern Territory, which has recently assumed the standing of a state in this respect.

Each state government has established an education department that provides, without fees, primary and secondary schooling within that state. The federal or commonwealth government has direct responsibility for schooling only within the Australian Capital Territory. Consequently, the provision of secondary schooling and the conduct of public examinations is undertaken within eight separate systems of education, whose structure and functioning differ in significant ways. While the arrangements made for schooling at the secondary level and the conduct of public examinations are clearly similar between the states, interesting and important differences exist, particularly in certification at the final year of schooling and in selection for entry into institutions of higher education.

Within each state a further distinction can be made between government schools, which are financed and staffed by the state education departments, and the non-government schools, which receive a limited amount of aid from the state and federal governments. The non-government schools, in the main, have been established by religious organizations. In particular, the Catholic church has an extensive system of secondary schools, the majority of which are operated by religious orders. The Protestant churches also operate secondary schools, many of which tend to be prestigious and expensive. Government schools cater to approximately 75 percent of all secondary students. Of the remainder about two-thirds are in Catholic schools. These non-government secondary schools tend to be single-sex, academically oriented institutions, which in general charge substantial fees and cater to children from upper- and middle-income families. The government schools are free, secular, and in the main comprehensive, coeducational schools that provide a general education for students up to the age of sixteen. Victoria, however, also has a system of technical schools which provide a secondary education with a vocational and more practical emphasis.

One of the consequences of the provision for education in both government and non-government schools is that the public examinations are conducted not by the state education departments but by independent statutory authorities or boards in all but three state systems. All types of secondary schools are represented on these boards as well as institutions of higher education, particularly universities. The universities formerly conducted matriculation examinations for the selection of their students. However, as the credentialing function of the examination at the final year of schooling has become of greater importance, the universities with some reluctance have passed over all responsibility for the conduct of public examinations to the more broadly representative boards.

Education in Australia is compulsory for all children between the ages of six and fifteen years, although different regulations for commencing and leaving school apply in each state. The minimum leaving age is fifteen years in all states except Tasmania, where it is officially sixteen years. Three states and the Northern Territory provide for seven years of primary schooling and five years of secondary schooling. The remaining states and the Australian Capital Territory provide for six years of primary and six years of secondary schooling. All students progress automatically from primary to secondary schooling. No selection procedures exist at this point, and the public examinations that used to exist at this level were abolished about forty to fifty years ago. Thus, secondary schooling is virtually universal up to the age of fifteen years. However, an increasing proportion of the age group is continuing with secondary education beyond this minimum leaving age. The size of the age cohort in Australia is approximately 250,000 students, and of these in 1983 94 percent were enrolled in year 10, 64 percent in year 11, and 41 percent in year 12.

Changing Retention Rates

Retention rates at year 11 and year 12 in secondary schools vary between states and territories according to the demographic composition of the population, the organization of the school systems, and traditions developed over time. While the degree of urbanization and the employment opportunities for youth also influence retention rates, it has become increasingly apparent during recent years that the nature of the courses offered by schools during years 11 and 12 and the certification and examination procedures have significant effects. Nevertheless, it is clear that students in the final years of secondary schooling are drawn disproportionately from families of professional and managerial occupations. As a corollary, students from lower socioeconomic backgrounds are not represented in school programs at years 11 and 12 in the proportions in which they are present in the community. Furthermore, in 1983 approximately 93 percent of students in the non-Catholic, non-government schools completed twelve years of schooling, compared with 51 percent in the Catholic secondary schools, and 34 percent in the government schools. These figures reflect the differences in the status backgrounds and in the levels of aspiration of the students attending these different types of schools.

Retention rates rose from 23 percent in 1963 to 33 percent in 1973, after which they levelled off for nearly a decade; they rose in 1983 to 41 percent. These figures mask a continued rise during the last decade in the retention rates for girls and a fall in the retention rates for boys. The proportion of girls surviving to year 12 now exceeds by approximately 5 percent the proportion of boys. These increases in retention rates resulted from measures taken during recent years. Financial assistance has been extended to poor students, including aboriginal students, to assist them to complete secondary schooling. In addition, changes have been made to enable schooling to become more relevant to the needs of disadvantaged students, and attempts have been made to integrate the schools more closely with their communities. In 1984 the Commonwealth Schools Commission introduced a participation and equity program in order to increase further retention rates, particularly among disadvantaged youth (Australia, Commonwealth Schools Commission, 1984).

Awarding School-Leaving Certificates

Until the mid-1960s each state operated a two-level system of public examinations. After ten years of schooling an external statewide examination was conducted. Those students who were successful in this examination (which measured achievement in school subjects having fixed courses of instruction) were issued a certificate that had statewide currency for the purpose of gaining entry to employment, apprenticeships, and other training. Students who gained the first certificate and continued at school until year 12 took a matriculation examination. This examination was conducted or strongly influenced by the state universities and provided the basis for selection for entry into tertiary institutions. Thus, each examination served a specific purpose: the year-10 examination was for certification, and the year-12 examination largely was for selection into tertiary education.

Three factors have operated to substantially change this picture. One was the increased range and diversity of tertiary institutions. New universities, colleges of advanced education, and colleges of technical and further education were created in the educational expansion of the 1960s and early 1970s. Each of these institutions has different policies for entry. The former matriculation requirements for entry into the single university in each state could no longer justifiably be imposed on all students seeking admission to other post-secondary institutions and indeed to the newer universities established within five of the six states. A second factor was the increased proportion of the student population staying on to complete the final year of secondary schooling, many of whom were not interested in, nor suited to, continuing with university studies. These students required some form of certification to indicate their level of educational achievement for the purposes of gaining employment, but they did not need the information for selection into higher education. A third factor was the growing reaction by teachers against the dominance of university-entry requirements on the curricula of the secondary schools. Teachers claimed that their options to teach what they thought appropriate and the options of their students to study what they considered relevant were being unnecessarily restricted. Many students were being forced into the study of subjects for which they were not suited because of the requirements associated with selection for a university education and for entry into the most prestigious courses.

The first real change in the certification procedures and in the public examination systems of the Australian states was the internalization of examinations at the year-10 level. Each state has replaced certification based on an external statewide examination at this level with an assessment program undertaken within the school, generally without formal certification. This has left schools free to develop their own curricula at the lower secondary level and up to the end of year 10, since all external examinations other than those in the final year of schooling have been abolished. However, as a consequence, the examination and certifications systems at the year-12 level must serve the dual function of selecting for entry into higher education and certifying for the successful completion of a full period of twelve years of schooling.

Each of the eight examination systems in the states and territories proceeded in a different way to meet the changing demands placed on them. Two major sets of problems arose. The first is associated with providing opportunities for schools and teachers to develop their own courses at the year-11 and year-12 levels. These problems are concerned with the acceptance of new courses by students, parents, employers, and the tertiary institutions; and with courses which might still be provided by the examining boards. The second set of problems is associated with permitting teachers to provide assessments of student performance, either as the total indicator or as a partial indicator of achievement. In either case, if equivalence is to be established between assessments derived from different sources, then some form of moderation between the different assessments is required. A further problem relating to the use of teacher assessments is whether it is possible to shift from the use of assessments which are normative in nature (comparing students and groups of students) to assessments which are criterion-referenced (in which some attempt is made to measure student performance against an absolute standard).

Many changes have occurred in the operation of the examination systems at the year-12 level during the past decade, and further changes have been planned. The first major change is the diminution of the influence and importance of the older state universities in the determination of the curricula and the conduct of the matriculation examinations at the terminal secondary-school level. Secondly, partly as a consequence, schools, teachers, and the students themselves in some instances have a greater say in the development of appropriate courses. Thirdly, greater recognition is given to the judgment of teachers in assessing student performance. Finally, in some cases, deliberate efforts are made to separate the certification or credentialing function of the year-12 examinations from their selection function, which involve providing information to tertiary institutions to assist with the selections for entry into these institutions. The developments in each state will be considered in turn. They show a wide range of possible responses to these complex problems.

The Australian Capital Territory. The system that has proceeded furthest in introducing change is in the Australian Capital Territory (ACT). In 1974 responsibility for schooling in the government schools within the ACT was transferred from the New South Wales Education Department to the ACT Schools Authority. A new and independent system of assessment and accreditation became necessary. The ACT Schools Accrediting Agency was established within the ACT Schools Authority to service both government and non-government schools. The main feature of the procedures set up by the agency is the requirement that the teachers in each school develop their own courses of instruction at the year-11 and year-12 levels, submitting some, but not all courses, for accreditation. Students can take three types of courses: (1) registered courses, which were developed to meet the cultural and recreational needs of students and were not submitted for accreditation but merely registered with the agency; (2) accredited courses, which were examined by the agency for educational soundness for study at years 11 and 12, and, where appropriate, accredited; and (3) T-classified courses, which were examined by representatives of the Australian National University as to whether they provided appropriate preparation for tertiary study.

Information on the performance of students in these courses is provided in two parts: a secondary college record and, for students wishing to proceed to higher education, supplementary information for tertiary entrance. Providing the information in two parts separates the certification function from the selection function of the Australian Capital Territory Year 12 Certificate. All assessment is internal to the school. Attempts are made to establish procedures to ensure comparability of standards both between and within schools of the assessments awarded in particular subject areas. Considerable flexibility is provided for setting up requirements for students to select appropriate programs of courses, either for general education or for preparation for entry into higher education. Furthermore an alternative school, the school without walls, was established with an integrated program that was accredited so that assessments of student performance could be made and reported. This flexibility has enabled the retention rates within the ACT to rise markedly during recent years, and students have expressed a greater degree of satisfaction with the system than was previously experienced (Anderson, Saltet and Vervoorn, 198[illegible] Nevertheless, in spite of the variety and flexibility in the courses offered, there is a marked tendency for students to select courses and a program to study that prepare them for entry into institutions of higher education.

New South Wales. The system that remains closest to the traditional mold is that of New South Wales. The Board of Senior School Studies is dominated by the New South Wales Department of Education, and the board's centrally determined syllabuses and examinations attract a great majority of the students in the year-11 and year-12 levels. It is also possible for individual schools to develop courses referred to as "other approved studies," which are submitted to the board for approval as alternatives to the traditional subjects which the board sponsors. Once approval for such subjects is obtained, a simple assessment of performance is recorded on the certification which is issued by the board.

In general, a student's performance in a subject is determined partly (50 percent) by achievement on a public examination and partly (50 percent) by an estimate made by the teachers of the expected level of performance of the student on the examination. The estimated mark is scaled by moderation procedures across schools using the level of school performance on the examination. Assessments made during the two-year program are not used directly in the assignment of marks.

The marks recorded on the certificate to indicate the level of performance of students are rescaled to take into account differences in the quality of the candidates in different subjects. To do this the marks awarded in all subjects to each student on the public examination are added to the marks assigned as moderated school estimates of performance. This provides a measure of the student's total level of achievement. These measures are then assigned to each student to obtain an indicator of the level of achievement of the candidate taking each subject. Thus, the performance of the subject candidate is adjusted to take account of differences between candidates of different subjects. This process is repeated through several iterations until stability in the adjustments occurred. In this way rescaling is undertaken to yield a performance measure, equivalent to that which would have been obtained if all students taking the public examinations had taken that particular subject.

Unfortunately, this complex procedure is not readily understood by examiners, students, and teachers and is subject to much criticism (Keeves and Parkyn, 1980). Unfortunately, too, the procedure introduces a bias that favors high-performance students who take mathematics and physical sciences subjects that correlate highly together and can be marked with a high degree of reliability. At the same time this rescaling procedure is disadvantageous to students who perform at a low level in the same subjects. Since more males than females sit for these subjects, a substantial degree of sex bias is introduced. This is demonstrated by the very high proportion of male students who perform at the highest levels in the examination.

A proposal for a change in this system has been made. The Board of Senior School Studies advocates that the use of teachers' assessments of performance replace the teachers' estimates. The public examinations will be retained and the moderation of teachers' assessments using examination performance will continue. The rescaling of marks between subjects, however, will be abandoned. In addition, Swan and McKinnon (1984) have proposed broadening the range of subjects from which year-11 and year-12 students might choose and providing a greater role for schools in developing the subjects to be studied by their students.

Victoria. In Victoria, responsibility for terminal secondary-school examinations and certification was transferred in 1978 to the Victorian Institute of Secondary Education (VISE), which is less dominated by the universities and more representative of the secondary schools themselves. VISE divides subjects into two groups. For subjects in group 1, VISE prescribes syllabuses which have included a core and optional components. The core is assessed by an external examination, which counts at least 50 percent and customarily 70 percent of the total mark; the optional components are assessed by the school and moderated between schools using performance on the external examination. Allowance is also made for differences between candidates for each subject by using an iterative rescaling procedure similar to that employed in New South Wales. However, performance is reported on the Higher School Certificate in the form of letter grades on a six-point scale, as well as through the use of rescaled marks.

Group 2 subjects are devised by schools, although they have to be accredited by VISE. These subjects are totally assessed within schools. They are reported either by six-point letter grades or on a two-point letter scale: S for satisfactorily completed; N for not satisfactorily completed. An alternative School Year 12 and Tertiary Entrance Certificate (STC) is accredited by VISE as a group 2, higher-school certificate course. In this course, students negotiate an appropriate program of work with their school. In 1983 twenty-eight schools were providing this course alongside a traditional higher-school certificate program. For 1985 the number of schools offering this course was expected to increase to seventy-six. In addition, a similar technical year-12 course, in which details of the course are negotiated between students and their teachers, was introduced in the technical schools. In colleges of technical and further education, courses which are referred to as tertiary orientation-program courses were developed and, in 1984, they attracted very substantial numbers, estimated to be eight thousand students.

VISE has sought public reaction to plans for a very substantial revision to the higher school certificate examinations. A ministerial review of post-compulsory schooling (1984, 1985) has indicated its support for change and has proposed the development of a single certificate which would be used across educational institutions of all types and for courses of all types at the end of twelve years of education. The variety in the provision for education at the pre-tertiary level and the radical nature of many of the proposals being advanced pose particular problems for the Victorian universities and colleges of advanced education.

Queensland. As the result of an investigation into the state's examination system in 1970, Queensland abolished external examination for year-12 certification completely. A Board of Secondary School Studies was set up as a statutory authority to issue certificates and approve syllabuses developed by subject advisory committees. In addition, individual schools can develop subjects which are registered by the board, but which do not count for tertiary entrance. However, schools have not very extensively used their right to develop their own courses.

The schools provide the assessment of student achievement in each subject. The board exercises its responsibility for maintaining comparability of grading partly through a system of moderation based on area meetings of teachers (in which they discuss the standards of courses and compare students' work) and partly by a system of regional and state review panels and a program of reference testing. Reviews of this system identified several difficulties. Subsequently, student achievement was assessed against defined criteria of performance rather than by comparing students one with another (Scott et al., 1978). Assessments of student performance against specific criteria is reported in terms of five categories of level of achievement. This is the first attempt at the terminal secondary-school stage in Australia to reject normative comparisons and to endeavor to employ a criterion-reference approach. The success of this system, which was introduced in the early 1970s, is reflected in a striking rise in retention rates in Queensland schools at both the year-11 and year-12 levels.

South Australia. In recent years two systems have been operating at the year-12 level in South Australian secondary schools. The Public Examinations Board issues certificates to students who have been successful in an external matriculation examination. Twenty-five percent of the total score for each subject is provided by school assessments which are moderated by the students' scores on the examination itself.

While student performance is reported on a six-point numerical scale, the assignment of categories takes place after the students' examination marks and school assessments are rescaled using an iterative procedure similar to that employed in Victoria and New South Wales. The problems encountered in these two systems also exist in South Australia.

A second system has been set up by the South Australian Education Department which provides alternative courses for both government and non-government schools. Student performance for the Secondary School Certificate associated with these courses is reported on a five-point letter scale. Moderation takes place between schools to achieve some comparability of assessments either through visits of moderators to schools or through meetings of teachers at which samples of students' work and the corresponding assessments provided by teachers are examined. Two committees of inquiry have considered reforms which might be introduced (Jones, 1978; South Australia, Committee of Enquiry into Education, 1981), and a Senior Secondary Assessment Board of South Australia replaced the Public Examinations Board in early 1984. The new board maintains the two systems of subjects, but all results are reported on a single certificate. More extensive changes are likely to follow.

Western Australia. Until 1975 Western Australia had an external examination conducted at the year-12 stage at both the leaving and matriculation levels by the University of Western Australia. Since that time a Board of Secondary Education has been established, which acts alongside a Tertiary Admissions Examination Committee. Two categories of subjects are offered at years 11 and

12. Certification of secondary education subjects, which are general in nature, are offered by the Board of Secondary Education, and Tertiary Admissions Examination subjects are offered jointly by the board and the Tertiary Admissions Examination Committee. The board issues a certificate recording results for both types of subjects. For the board subjects, various approaches to moderation are used in order to achieve comparability of assessments between schools. In some subjects comparability of assessments is obtained through visits of moderators to the schools or through meetings of teachers that review students' work and teachers' assessments. In other subjects, moderation tests are given and school assessments are adjusted to match the score distributions on the reference tests. For the Tertiary Admissions Examination subjects an equal weighting of examination marks and school assessments, moderated between schools using examination marks, is used. On the Certificate of Secondary Education a students' performance is reported in terms of decile ranks within subjects.

The Certificate of Secondary Education also records an index of academic standing, which is calculated by the Board of Secondary Education using group performance on the Australian Scholastic Aptitude Test (ASAT) to moderate marks in different subjects for differences in the level of difficulty of the different subjects which arise from the differences in ability of the candidatures attracted to the different subjects.

Following a review of these examination and certification procedures (McGaw, 1984), changes were adopted to take effect at the year-11 level in 1985. A single new agency, the Secondary Education Authority, was established. A revised Certificate of Secondary Education is the only certificate issued to students. Two categories of subjects continue to operate: a smaller set associated with entry into tertiary institutions, and a second set of all other accredited subjects available as alternative courses of study. A student's achievement in different subjects during the last four years of schooling is recorded using a five-point letter scale, and achievement on year-12 subjects which count for entry to tertiary institutions is reported on a hundred-point scale. These marks are a combination of an equal weighting of external examination marks and school assessments moderated using the external examination marks.

Tasmania. The School Board of Tasmania awards a Higher School Certificate to school leavers based on performance on subjects taken at two different levels. Subjects at the lower level are assessed within the school. Moderation of assessments between schools is undertaken by subject advisers visiting schools and in some subjects, such as mathematics, are carried out through the use of a centralized item bank to maintain quality control on the assessments awarded by a school. Subjects at the upper level are assessed by means of external examinations together with school assessments which are moderated by the external examinations and which contribute between 25 and 50 percent to the final mark. Adopting criterion-referenced assessment and perhaps abandoning external examinations altogether have been discussed in Tasmania, and work was also undertaken during 1984 to develop criterion-referenced tests in language and mathematics. These would be given to all students at the year-6 (or end of primary school) level and at the year-10 (or end of compulsory schooling) level, and certificates of performance issued.

Northern Territory. Since 1984 students at year-11 and year-12 levels on completion of their schooling receive a Northern Territory Senior Secondary Studies Certificate, which is issued by the schools and records information on all subjects taken by the students. Students are able to take the subjects of the Public Examinations Board of South Australia and will be able to take those of the new Senior Secondary Assessment Board of South Australia in the future. Students completing any of these year-12 subjects are also issued the corresponding South Australian certificate as well as the Northern Territory certificate. It is also possible for students to take school-developed subjects which are suitable for study at year-11 and year-12 levels and which are accredited within the Northern Territory Education Department.

In the programs of nearly every one of these eight systems some attention is being given to changes that will increase the flexibility with which courses are developed within schools and accredited by the systems involved. In addition, changes were introduced in the ways in which school assessments are made and recorded, in a fair manner, on a certificate that is issued to the student on completion of schooling. An important development is the attempt made to break away from normative assessments and to use criterion-referenced assessments. However, this has given rise to some problems concerned with the degree of flexibility available to schools to develop their own courses.

Selection for Entry to Tertiary Education

Two significant problems associated with selection for higher education accompany the marked increases in opportunities for study at universities and colleges of education and in retention rates at the year-12 level. First, the greatly increased number of applicants for places in university courses requires the employment of admission procedures that are not only as equitable as possible but are also seen to be fair to those students who are not selected. Moreover, by tradition this selection takes place within the space of a very few weeks. The processes that are employed are relatively simple, clearly defined, and open to scrutiny. Therefore, reliance on the use of a total score obtained in a matriculation examination is heavy. However, some institutions of higher education use a profile of information, especially for borderline cases, which include, for example, performance on the Australian Scholastic Aptitude Test (ASAT). Of particular consequence in the selection process is the need for a high degree of selectivity for entry into the prestigious courses of medicine and law within the universities. The competition for places in these courses is great, and a level of accuracy is demanded in the selection process that is unwarranted and probably impossible to provide under the conditions which normally apply in the conduct of public examinations.

The second problem which has arisen has also been a consequence of the increase in retention rates over the past twenty years. When only a small proportion of the age group remained at school to the year-12 level, it was necessary to provide only a limited range of courses in preparation for study at a university. Because the range of ability of the students taking these courses was restricted, the different courses attracted students of approximately equal average abilities. Thus it was relatively easy to set a pass mark at 50 out of 100 and simple to add the marks obtained on the examination papers in different subjects to produce a total score. However, with increased retention rates, students of lesser ability select some subjects in preference to others, on the grounds that they are less demanding and that a shorter period of preparation is required for the study of the subject at the year-12 level. The problem facing those concerned with the selection of students for entry into tertiary institutions is how to compensate for the differences in the quality of the candidates taking the different subjects or taking the same subject at different levels of complexity.

A related issue has also arisen as to whether the allowance made for such differences in the quality of subject candidates, which is necessary for the selection process, should be made clear on the certificate provided to the student on the completion of schooling at the year-12 stage. It is necessary to recognize that making these adjustments is essentially a requirement of the selection process and not of great consequence for the certification process, although the given level of difficulty of a subject studied could be of interest to those using the information contained on a certificate. In the sections that follow, the manner in which each of the systems provides information for selection into higher education, both to the tertiary institutions and to the student, is considered. As the Northern Territory does not have a full range of tertiary institutions, the majority of tertiary students from this system enter institutions in South Australia on the basis of results in subjects sponsored by the Senior Secondary Assessment Board of South Australia. Under these circumstances selection for entry into tertiary institutions in the Northern Territory is not being considered here.

Australian Capital Territory. The ACT Schools Accrediting Agency gives to both the student and to the tertiary institution what has been referred to as supplementary information for tertiary entrance. The information provided includes a course score, on a scale with a mean of 65 and a standard deviation of 15 as awarded by the school that the student attended, as well as an adjusted course score to allow for differences in candidates between courses and schools for each T-classified course taken by a student. The adjusted scores are added to obtain an aggregate tertiary entrance score for each student by taking the student's three best scores on major courses together with 0.6 of a fourth or minor-course score. In addition, percentile rankings within the groups of students for whom a tertiary entrance score was calculated, and also within the age cohort in the Australian Capital Territory, are provided.

The rescaling of course scores to obtain equivalence between subjects and between schools is undertaken using the mean scores of groups of students on the ASAT. The test is a three-hour, one-hundred-item, objective test designed to assess aptitude for study in the areas of the humanities, mathematics, the sciences and the social sciences. Because this test includes items associated with learning in the areas of mathematics and the sciences, sometimes substantial differences result in the mean scores of male and female students on ASAT. A research investigation into sex differences in performance on ASAT (Adams, 1984) has indicated that differences between the sexes in their performance was due to differences in the retention rates for male and female students at the year-12 level, to differences in the extent to which males and females studied mathematics and science, and to differences between the sexes in their confidence in success on ASAT. It was also found that after allowance had been made for these factors, little evidence of bias in performance due to the socioeconomic backgrounds of the students emerged. In 1983 a small correction to increase the scores of girls was considered necessary to adjust for differences between the sexes in performance on ASAT. However, largely because of changes in retention rates, with an increase proportion of boys remaining at school at year 12, compensation was not required in 1984.

New South Wales. An aggregate mark, based on adjusted scores for ten half-subjects, together with a percentile ranking for the student, obtained by using the aggregate score, is provided for tertiary selection. The scores are adjusted, as explained in the previous section, using an iterative procedure to provide for differences between the quality of the groups of students taking different subjects. The Board of Senior School Studies has proposed abandoning the scaling of subject marks which allows for differences in candidates, and it will be left to the tertiary institutions to undertake their own adjustments to subject marks in order to calculate an aggregate score similar to the one at present available. A difficulty could arise insofar as the sizes of the subject groups for many subjects would be too small for the iterative procedure to operate effectively, if the rescaling process were undertaken within each tertiary institution.

Victoria. Admission to tertiary institutions in Victoria is based largely on an aggregate score, which is calculated using performance on a student's best four group-1 subjects and 10 percent of the fifth group-1 subject. The aggregate score combines the subject scores, which are adjusted for ability differences between the groups of students taking different subjects through the use of an iterative procedure similar to that employed in New South Wales. However, in addition, some universities and colleges of advanced education accept high performance in group-2 subjects, in the school year 12 and Tertiary Entrance Certificate courses, and in the tertiary orientation program for entry to certain tertiary courses. Difficulties have arisen in establishing comparability for purposes of selection between the different sets of results recorded on the different certificates. These problems would be greatly accentuated if the current proposals, which have been made by the Victorian Institute of Secondary Education (1984) and the ministerial review of post-compulsory schooling (1984, 1985), were implemented. In part the criticisms advanced against the current system are concerned with the bias introduced in the iterative procedure that is associated with clear advantages to boys who have taken mathematics and physical science courses.

Queensland. For admission to tertiary institutions, the Board of Secondary School Studies issues students with a separate tertiary entrance statement which records the aggregate tertiary entrance score. The students are required to complete a specified program of work over a two-year period. Their school assessments in each course approved by the board are rescaled using performance on the ASAT to allow for differences between subjects. The rescaled scores are added to form a total score for each student within the school, and subsequently these total scores are rescaled again using ASAT to make allowance for the differences in student ability between schools. A rescaled aggregate score for each student is then calculated. These rescaled aggregate scores are used to rank all students in order of merit within the state, and a tertiary entrance score on a rectangular distribution with a maximum of 1,000 is obtained from the ranking. The lowest tertiary entrance score is set by the relative size of the eligible year-12 group for whom tertiary entrance scores are calculated with respect to the seventeen-year-old population. Thus, if the proportion of the tertiary entrance group is 35 percent of the seventeen-year-old population, the lowest score is set at 650. Selection for entry to tertiary institutions is based on the tertiary entrance scores. However, there is strong criticism within Queensland universities and schools that the rescaling procedures using ASAT do not give adequate recognition to the quality of teaching within a school.

South Australia. Entrance to universities is based solely on performance in the subjects of the Public Examinations Board. The colleges of advance education are willing to take into account performance on Secondary School Certificate subjects. Marks on the Public Examinations Board subjects are adjusted using a iterative procedure similar to that employed in New South Wales and Victoria. An aggregate score is calculated, recorded on the certificate issued by the board, and used in selection for entry into tertiary institutions. However, the reconstituted board proposes that the rescaling of scores and the calculation of an aggregate no longer be undertaken by the board. Whether the tertiary institutions have adequate numbers of students in all courses for them to be able to undertake the rescaling within the institutions to produce meaningful results seems unlikely.

Western Australia. Until 1985 selections into tertiary institutions was based on the aggregate score obtained using 10 percent of performance on the ASAT test and rescaled marks awarded on Tertiary Admissions Examination subjects, without use of any school assessment component. ASAT was used in the rescaling. This information was provided to the student on a supplementary certificate. In 1986 under the new scheme advanced by the McGaw Committee (1984), an admissions average score will be calculated provided that a student has achieved a satisfactory level of literacy and has recorded satisfactory performance in six subjects during the year-12 course, with five of these subjects being year-12 subjects. The admissions average score will be based on three or more subjects, with at least one being a quantitative/science subject and at least one a humanities/social studies subject. The averages on three, four, or five subjects will be calculated, subject to these requirements to obtain the best average score, which will be used for tertiary entrance purposes. All scores used for tertiary entrance will continue to be rescaled using ASAT, and a small component of a student's performance on ASAT will be included in the calculation of the admissions average score.

Tasmania. Until 1981 entry into the University of Tasmania required passes in four approved higher-level subjects. From 1982 onward it has been necessary for a student to also obtain a lower pass or better in two further higher-level subjects, which are either approved or non-approved subjects offered at this level, or a pass in two further lower-level subjects for which there are corresponding higher level subjects. It is possible for students to matriculate for entry to university from year 11. However, most students undertake two years of study beyond the end of year 10.

Special Issues of Certification and Selection

The Issue of Control

Perhaps the first significant issue is the lessening of the influence of the state universities on the process of certification. The changes that have taken place have had a significant influence on the conduct of the examinations and on the curricula of the upper secondary schools. In five states statutory examining authorities have now been established, and control no longer rests with either the universities or the state departments of education. However, in the remaining three systems the examining authorities work within the state departments of education, and the director-general or the minister of education is thus directly responsible for their operation. In all states the examining authorities include representation from the education departments, the schools (whether government, religious, or independent), and the growing range of tertiary institutions. Responsibility for selection for higher education remains with each university and college of advanced education but is necessarily based on information provided by the examining authorities.

Responsibility for the Development of Courses

In the Australian Capital Territory, in particular, the responsibility for developing courses has devolved to the schools themselves, and in each of the other states increasing opportunities for schools to develop their own courses exist. However, in many states the currency of these courses both for selection for entry into tertiary institutions and for certification and use in gaining employment remains uncertain. While there is growing acceptance that courses developed within schools should be recorded on the certificate alongside courses developed by the examination boards, it is unclear as to whether such courses should be identified in such a way on the certificate that employers and parents readily recognize that the courses might be of lower standing. It is possible too, that if such courses do not lead to entry into tertiary institutions, they will be recognized by both teachers and students as inferior.

Internal Awareness

From this summary of terminal secondary school examinations in Australia, it is evident that the movement is away from state-wide external examinations toward internal assessment of student performance. The Australian Capital Territory model is perhaps the most developed, with standards being maintained by a system of accreditation and moderation. However, what is within a compact and small region, such as the Australian Capital Territory, might not be easily achieved in a state in which the population is widely dispersed. Yet in Queensland, considerable movement has taken place along similar lines, indicating that problems of distance can be overcome. It is clear that developments away from centralized courses and external examinations toward the accreditation of courses and internal assessment are accompanied by the devolution of authority. This is achieved by the examining authorities becoming less autocratic and adopting a more supportive and facilitating role in accrediting courses, moderating assessments, and issuing certificates on the advice of the schools concerned.

The Certificates

One of the moves which has assisted these developments is separating the certificate into two parts. One part serves a certification function associated with the completion of a full period of secondary education to the end of year 12, and the other part, generally less substantial in format, serves as a record of the information available for use in selection for entry into tertiary institutions. However, the information associated with tertiary selection has generally assumed such importance that students, parents, and employers attach greater weight to this document than to the information on secondary schooling completed. As a consequence considerable pressure has been exerted to further reduce the information made available for tertiary selection, even to the extent of suggesting that detailed information, in the form of marks gained in school assessments and examinations, no longer be provided to tertiary institutions.

Possible Bias in the Information Provided

Questioning the information provided to the tertiary institutions has come, in the main, from the possibility of bias arising in the procedures employed for scaling school assessments and scaling external examination marks. These scaling procedures seek to take into consideration the differences in the quality or ability of the groups of candidates coming from different schools or taking different subjects. The iterative procedure appears to give an advantage to boys, who tend to take mathematics and the physical science subjects. The marking processes in these subjects have greater reliability, and the scores on these subjects, greater correlation. At the same time that the better performing students in these subjects have possibly gained an advantage, the weaker students have been disadvantaged. However, this is largely overlooked since it is only the better students who are rewarded by entry into the more prestigious courses. Likewise the nature of the ASAT is possibly advantageous to those students who have studied mathematics and the physical sciences, particularly because boys congregate into those groups which have benefited in the rescaling operation.

The Use of the Australian Scholastic Aptitude Test (ASAT)

Some of the criticisms of the use of the ASAT test are ill-founded. Such criticisms are based on only moderate correlations in some subjects, at a level of analysis between students, or between performances on the ASAT and teacher assessments or performance in examinations. The relevant correlations in the rescaling operation are those associated with groups of students and the information available on such correlations has shown that they are higher than generally expected. Their magnitudes largely depend on the extent to which students of like ability have clustered together into subject and school groups. What is lacking is widespread research on the predictive power of the aggregate scores used in selection for success at the tertiary level. Only the universities and colleges of advanced education can provide this evidence. However, the evidence that is available in published form (McGaw, Warry, and McBryde, 1975; McGaw, 1977; and Müller, 1982) indicates that aggregate scores obtained from school assessments correlate as well with university performance as do aggregate scores obtained from external examinations.

The Future Use of Reference Tests

A very substantial problem is likely to arise if procedures are established by which universities are not provided with the information they require to select students for entry as fairly and as accurately as possible, particularly into those courses for which keen competition exists. Universities face serious problems in this regard which have arisen from the diversification of programs that lead to entry into tertiary education. In recent years, less than half the applicants to some universities have submitted the results of their performance on an external examination taken during the previous year at the terminal stage of secondary schooling. In order to alleviate these problems, and to have a tried-and-tested alternative in the event appropriate information were no longer available, some institutions in Victoria have investigated the use of reference tests in mathematics and language for rescaling assessments obtained from a variety of sources and for defining threshold levels for entry into tertiary institutions (Beswick et al., 1984).

Item Banks

Masters (1985) has also proposed the use of items banks which are calibrated using the logistic function to provide a reference scale for the rescaling of school assessments in a normative-referenced mode and for obtaining a performance profile in a limited subject field in a criterion-referenced mode. The use of an item-banking approach has been tried in Tasmania in recent years. However, the computer-based hardware and software for the efficient operation of this approach are only now becoming widely available. Moreover, the use of this approach requires a significant body of teachers with a thorough understanding of the technical aspects of the procedures employed to have full confidence in them.

A Profile of a Total Score

One of the proposals that has received some acceptance in several states is that tertiary institutions no longer be provided with a total score that could be used for selection into tertiary courses. There are three reasons for this proposed change: first, the procedures used in the calculation of a total score cannot readily be explained to students and teachers; second, possibility of bias exists in the procedures employed in rescaling that favors students studying mathematics and the physical sciences, and, thus, male students in preference to female students; and third, the generally recognized desirability of universities, using a profile of information rather than a single aggregate score on the selection process.

However, it is important to recognize that universities have sought to select students in ways that are fair and equitable and free from preferential treatment for individual students or groups of students. Moreover, a total score based on examination performance has been shown to be the best single predictor of academic success in higher education. Consequently the universities seek to continue using a total score in the selection process. If a total score were no longer provided, alternative selection procedures would need to be developed based on a profile of information and a series of decision rules. These would be developed from research studies into the predictive validity of the different components of the profile for performance in tertiary studies. Likewise, it would be possible to develop a weighted composite from the components of the profile for use in the selection process. This weighted composite could also include loadings given for the study of specific subjects at the year-12 level that are considered relevant for study during higher education.

Increases in Participation

One of the major problems facing Australian education is the need to increase the general level of participation at years 11 and 12. Ainly, Batten, and Miller (1984, a and b) have shown that diversification of the curriculum during these years of schooling can contribute significantly to increasing retention rates. However, the provision of financial support to students, the availability of employment opportunities, apprenticeships, and places in colleges of technical and further education have also been shown to influence retention rates. Retention rates are also affected by such characteristics as the students' sex, the socioeconomic status of their families, and the levels of aspiration of their families and peer groups. Diversification of the curriculum at the upper secondary-school level demands as a corollary greater flexibility in the selection procedures used for entry into higher education. The curriculum development center has circulated a discussion paper which considers some of the issues associated with certification in upper secondary education arising from increased diversification of the curriculum (McGaw and Hannan, 1985).

Academic Excellence

Some Australian universities contended that, in the best interest of both society and the individuals concerned, they must reaffirm their commitment to the goal of academic excellence both in courses undertaken in preparation for entry into universities and in the procedures that they employed to select those who should enter universities (Beswick et al., 1984: 89-90). It is also clear that the selection procedures cannot be divorced from the courses taken in preparation for entry to university study, since some of the courses taken at the upper secondary-school level are courses for further study, particularly in fields where knowledge and understanding are cumulative. For students who wish to proceed to university studies, taking courses that are not effective preparation for tertiary study squanders the final years of schooling and wastes the initial year at a university. Furthermore, all the evidence available seems to indicate that the strongest basis for selection for further education is performance in a foundation course at a lower level. The advantages of an external examination are that both an appropriate foundation course is defined and the level of achievement of students in that course is measured in a fair and reliable way. However, some important outcomes associated with the study of some courses cannot be validly measured by an external examination. Consequently, school assessments can make a valuable contribution.

Equitable Opportunities

The other major problem confronting Australian education is the apparent lack of equity in selection for entry to tertiary education between different social groups. The failure of Australian society to provide equal opportunities for all social groups at the stage of transition from secondary to higher education is well recognized. However, several research studies (Beswick, 1979; Blandy and Goldsworthy, 1975; and Lewis, 1977) have shown that the social selection that existed, took place during the upper secondary-school years. It is accepted that the universities should attempt to rectify any imbalance that they can, possibly by deferring selection for entry into the most prestigious courses at the tertiary level until after an initial year of study at an institution of higher education. However, it is also necessary for schools to provide appropriate programs of guidance and counseling so that all students who have the capacity to undertake higher education have the opportunity to do so.

Freedom of Choice for Students

Students need opportunities to exercise self-management, self-development, self-motivation, and self-discipline. This may be best achieved by providing a broad curriculum at the upper secondary-school level, with courses ranging from the highly abstract to the practical and applied. Students should then have the opportunity to make an informed choice based on sound guidance provided by the school. However, a difficulty arises insofar as many high schools throughout Australia are too small to provide a wide range of courses. The establishment of senior colleges would allow students greater freedom as well as provide them with a broad range of courses. However, the teacher unions in Australia strongly oppose proposals to set up senior colleges.

Comparison of State Systems

During recent years there has been increased mobility within the Australian states and territories. As a consequence students have transferred from one state system to another between the completion of the year-12 examination and entry into higher education. This has required institutions of higher education to develop procedures for equating performance in one state with that in another. This is achieved by the use of a model using the truncated normal distribution that was developed by Walker (1967) to account for the effects of selection. Using a procedure similar to that employed in Queensland for the calculation of rankings, allowance is made for the different levels of educational retention to the year-12 examination in the different states. As increasing proportions of an age cohort remain at school, and as greater variability occurs in the range of subjects taken which are considered appropriate for tertiary entrance, it will become increasingly difficult to apply this procedure for comparing the effects of differences in holding power between the state systems.

A Time of Change

The procedures of public examining in the eight Australian educational systems are in a state of flux. In developing appropriate solutions it is necessary to recognize that the debate is not so much about the procedures themselves but about their effects on teaching and learning at the year-11 and -12 levels. The critical issues are associated with the curricula of the upper secondary school which must now serve a range of purposes. At this level of education, students have prepared in different degrees for the work to be done, have different capacities to operate with abstract ideas or on concrete and practical tasks, have different powers to express themselves, have different emotional drives to succeed in scholarly work, and have different long-term aspirations for their future employment and careers. Moreover, some of these differences have been present since birth, and others have developed significantly during the teenage years. As a consequence, educational institutions, particularly following the marked increase in retention rates that has occurred during the past twenty years, must provide a greater diversity of offerings in order to meet a greater range of student needs than they have previously encountered. In addition, it is necessary that the more able be fully extended by courses that demand excellence and that provide an intellectual challenge. It is necessary that new courses be provided for living and working in a society that will be greatly influenced by rapid technological and social change. It is also necessary that all courses prepare for further learning, not only in the immediate future at the tertiary level, but continuing throughout life. However, it is essential that courses are seen to be relevant to the present and future lives of the students taking them. The changes that must be made in public examinations in Australia to meet changing demands as well as to effectively serve the needs of both individual youth and the larger Australian society are still unclear.

References

Adams, R. J. Sex Bias in ASAT? ACER Research Monograph No. 24. Hawthorn, Victoria: Australian Research Council for Educational Research (ACER), 1984.

Ainley, J.G., Batten, M., and Miller, H. Staying at High School in Victoria. ACER Research Monograph No. 23. Hawthorn, Victoria: Australian Research Council for Educational Research (ACER), 1984a.

Ainley, J.G., Batten, M., and Miller, H. Patterns of Retention in Australian Government Schools. ACER Research Monograph No. 27. Hawthorn, Victoria: Australian Research Council for Educational Research (ACER), 1984b.

Anderson, D.S., Saltet, M., and Vervoorn, A. Schools to Grow In: An Evaluation of Secondary Colleges. Canberra: Australia National University Press, 1980.

Australia, Commonwealth Schools Commission. Participation and Equity in Australian Schools. Canberra: Commonwealth Schools Commission, 1984.

Beswick, D.G., Schofield, H., Meek, L., and Masters, G. Selective Admissions Under Pressure. Melbourne: Centre for the Study of Higher Education, University of Melbourne, 1984.

Blandy, R., and Goldsworthy, A. "Equal Opportunity in South Australia" Bedford Park, South Australia: Flinders University (mimeo), 1975.

Jones, A.W. (Chairman, Committee of Enquiry into Year 12 Examination in South Australia). Report of the Committee. Adelaide: Government Printer, 1978.

Keeves, J.P., and Parkyn, G.W. The Higher School Certificate Examination in New South Wales: The Report of the Review Panel to the Board of Senior School Studies. Sydney: Board of Senior School Studies, 1980.

Lewis, R. Social Class Bias in Transition from School to Study in Tertiary Institutions. VIER Bulletin 39 (1977):25-30.

Masters, G.N. Item Banks and Year 12 Assessment. Melbourne: Centre for the Study of Higher Education, 1985.

McGaw, B. The Use of Rescaled Teacher Assessments in the Admission of Students to Tertiary Study. Australian Journal of Education 21 (1977):209-225.

McGaw, B. (Chairperson of the Ministerial Working Party of School Certification and Tertiary Admissions Procedures). Assessment in the Upper Secondary School in Western Australia. Perth: Western Australian Government Printer, 1984.

McGaw, B., and Hannan, W. Certification in Upper Secondary Education. Canberra: Curriculum Development Centre, 1985.

McGaw, B., Warry, R., and McBryde, B. "Validation of Aptitude Measures for the Rescaling of School Assessment." Education Research and Perspectives 2 (1975):20-34.

Müller, W.J. "Prediction of Student Achievement at ANU from ACT College Assessments." Unpublished Masters Thesis, Australian National University, 1982.

Scott, E., Berkeley, G.F., Howell, M.A., Schuntner, L.T., Walker, R.F., and Winkle, L. A Review of School-based Assessment in Queensland Secondary Schools. Brisbane: Board of Secondary School Studies, 1978.

South Australia, Committee of Enquiry into Education in South Australia (Chairman: J.P. Keeves). Education and Change in South Australia: First Report. Adelaide: Education Department, 1981.

Swan, D., and McKinnon, K. Future Directions for Secondary Education: A Report. Sydney: New South Wales Department of Education, 1984.

Victoria, Ministerial Review of Post-Compulsory Schooling. Discussion Paper. Melbourne: Victorian Government Printer, 1984.

Victoria, Ministerial Review of Post-Compulsory Schooling. Report, Volume 1. Melbourne: Victorian Government Printer, 1985.

Victorian Institute of Secondary Education. Towards a Revised Policy on Curriculum and Assessment in the Victorian Year 12 HSC Program: A Paper for Discussion. Melbourne: Victorian Institute of Secondary Education, 1984.

Walker, D.A. "An Attempt to Construct a Model of the Effects of Selection." International Study of Achievement in Mathematics, Edited by T. Husen, vol. 2. New York: John Wiley and Sons, and Stockholm, Almqvist and Wiksell, 1967.

5

EDUCATION IN SWEDEN: ASSESSMENT OF STUDENT ACHIEVEMENT AND SELECTION FOR HIGHER EDUCATION

Sixten Marklund

Introduction

Sweden's population is 8.4 million. During the last fifty years the birth rate has been low, but this has been counteracted by an increased average life span and recently increased immigration.

The population was ethnically homogeneous until the 1960s. Small national minorities consist of the Lapps in the northern inland and a Finnish-speaking population along the northeastern boundary. During the last three decades immigration has exceeded emigration. Immigrants and their children now number one million; they live almost exclusively in the cities. Most immigrants have mother tongues other than Swedish, which has influenced the school system considerably. A new school subject for immigrants introduced in 1974 is Swedish as a Foreign Language. Nearly all immigrant students are also taught their home language at school. They are also given special study guidance and vocational orientation in their mother tongue. In 1974 a special "home-language teacher training" was also established.

Sweden's industrialization came late but developed quickly. The social structure of the population is roughly the same as in most industrialized countries. However, post-war policies on labor, salaries, and general welfare have considerably equalized the socio-economic classes. The differences in disposable income and working conditions seem to be smaller than in any other country.

The Overall School Structure

It is important to take into account rapid changes in the whole school system in recent years in considering the assessment of student results and the selection for higher education. After a nationwide experimental period from 1950 to 1962, the country moved from a basically dualistic school system, with different kinds of primary and secondary schools organized in a parallel structure to a unified comprehensive organization. This was fully implemented in 1972. Sweden now has nine-year, compulsory, comprehensive schooling for all from seven to sixteen years of age. This is followed by integrated, post-compulsory, upper-secondary schools throughout the country, comprising academic and general as well as technical-vocational education in the same organization.

The preschool structure was reorganized in 1967 and 1975; adult education, in 1967 and 1971; and higher education (universities and corresponding tertiary institutes), in 1975 and 1977.

This sequence of changes gave Sweden a totally new overall structure of education. The traditional stages of preschools, primary schools, secondary schools, and vocational schools were replaced by a system of pre-compulsory, compulsory, and post-compulsory schools. (figure 5.1.).

The school year starts in late August and ends in mid-June. It encompasses forty weeks. Five days of the school year are designated by the local authorities as "study days"; during this period teachers meet for planning and in-service training.

Classes are small by international standards, seldom more than thirty students. The mean size in 1982-83 was 22.6 in the compulsory schools and 24.4 in the post-compulsory schools. Because of the many possibilities that now exist to work individually with small groups or with companion teachers, the student-teacher ratio is much lower, about fifteen in compulsory schools and thirteen in post-compulsory schools.

The nine-year, comprehensive, basic school is the same all over the country. As can be seen in figure 5.1. it has two three-year stages of primary education (junior 1-3, middle 4-6) and a third lower-secondary stage (upper 7-9). English is compulsory during grades 3-9. Grades 1-6 usually have class teachers. Grades 7-9 have specialized teachers. The students have a common course of study, with the exception of 10 to 15 percent of the time in grades 7-9, when they have elective subjects (either a second foreign language or extra courses in their general subjects). Immigrant students can take their home language as their elective subject. Regardless of electives, a complete nine-year school course qualifies a student for study in the upper-secondary, post-compulsory school, the so-called gymnasium.

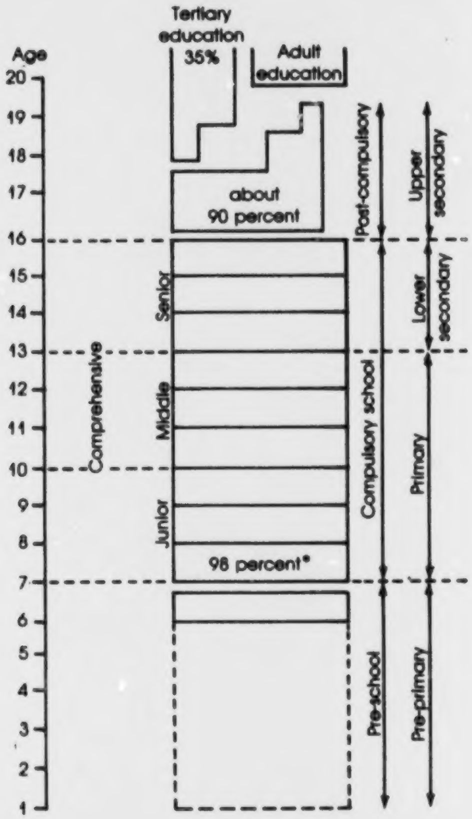
This integrated upper secondary school has three main sectors of study; one with art and social subjects, one with economics and commercial subjects, and one with scientific and technical subjects. Each one of these sectors offers academic, general, and vocational courses. These are of different length, from half-year courses to four-year courses (table 5.1.). The majority of students take one of twenty-two specifications. The rest (6 percent) take special courses for more narrowly restricted vocations.

Higher Education

For a long time, university policy required a studetexamen (secondary-school certificate) for admission to any of the faculties. In 1972 Parliament approved new principles for selecting students to higher education. According to these, every specialization in the upper-secondary school (in some cases after supplementary studies in Swedish and English), as well as other studies with equivalent aims and length, fulfills the general admission requirement for higher education. In addition, a person who is at least twenty-five-years old and has at least four years of occupational experience, regardless of schooling, also fulfills the general requirement. On top of this there are special admission requirements -- expressed as knowledge equivalent to full upper-secondary-school courses in special subjects -- for various programs and higher-education courses.

The new rules mean that, in principle, every adult is formally entitled to begin higher studies, and the expanded system of adult education provides opportunities for most people to acquire the prerequisite knowledge required for higher education. The rules for selection have proportional quotas for applicants from different backgrounds. In addition to school marks, working experience also entitles a person to qualification points.

Figure 5.1. School Structure in Sweden, 1984


Tertiary education
35%
Adult education
Age
20
19
18
17
16
15
14
13
12
11
10
9
8
7
6
5
4
3
2
1
about
90 percent
98 percent*
Comprehensive
Senior
Middle
Junior
Post-compulsory
Compulsory school
Pre-school
Upper secondary
Lower secondary
Primary
Pre-primary

*For every annual cohort, 1 percent go to special schools for physically and/or mentally handicapped, and 1 percent go to private schools (compulsory level). The other 98 percent join the regular basic school. Approximately 90 percent of the annual cohort continue to post-compulsory school and approximately 35 percent of the annual cohort go to universities and other kinds of post-secondary schools.

General Principles of Assessment

The Swedish school system is unitary in the sense that the same general and specific aims are pursued in the same kind of educational institutions all over the country. Thus, all those studying any given subject at the same level usually follow the same curriculum and have the same number of weekly periods.

The ideology underlying the structure of the school system is the principle that all students -- irrespective of social stratum or geographical location -- are to have the same opportunity of developing their personalities and acquiring knowledge and skills as far as their individual aptitudes and abilities will allow.

Courses and schedules are contained in a handbook, laroplan (referred to as the curriculum here), which states the overall aims of education as well as the aims and objectives of all subjects being taught, outlines the syllabus, gives guidelines for each subject, and discusses teaching methods and materials. The compulsory school and the upper-secondary school each have a curriculum. It should be noted that although guiding principles for teaching different subjects are laid down, it is clearly stated that within the general framework it is ultimately up to the teacher to develop the type of approach that best suits his or her personality. Single schools or groups of schools also run experimental programs.

Certain general aims are to be pursued, regardless of what subject is being taught. Among the general aims emphasized throughout school are clarity and order of thought, the ability to think critically and independently, the ability to resist tendentious influence, the ability to analyze and deal rationally with a problem, as well as the ability and willingness to cooperate with others.

Specific aims have also been set for each subject. They are mostly stated in rather general terms, but the curriculum also indicates well-defined study areas and activities that are to be included in the learning process. The evaluation of skills and knowledge has to be based on an analysis and interpretation of the aims laid down in the curriculum.

The outcome of such an analysis may vary owing to the subjective judgment of the person or persons making it. Consequently, the teaching of the same subject may differ to a certain extent, according to teachers' opinions on what are the most important aspects. It is also essential to make room for variation and change in order to adapt classroom work to the students' individual interests and aptitudes. Naturally, the freedom thus granted to teachers and students is limited by the necessity of conducting work in such a way that the stated aims can be attained.

When it comes to a large-scale assessment of achievements in any given subject, which requires the construction of special test batteries, it is imperative to reduce the subjective element in the interpretation of aims to a minimum. For this reason, decisions on the nature and extent of such an evaluation are always based on the consensus of a fairly large group of experts.

The Marking System

Marks indicating individual study results in a subject are given at the end of a term or academic year. A five-point scale is used, the figure 5 denoting the best achievement. Each mark is the expression of the teacher's overall assessment based on careful consideration of all relevant aspects. The individual teacher is solely responsible for the marking. No educational or legal authority can alter a given mark or force a teacher to do so.

Marks are given for the obvious purpose of telling students, and also their parents or guardians, to what extent their work during a certain period of study has been judged as successful. If this were the only function of marks, alternative channels of information could be used and the marking system could be abolished, as has indeed been suggested. This would eliminate or at least greatly diminish the element of competition which many students experience as a psychological strain.

However, marks have so far proved to be the most reliable instruments for the selection of students for admission to certain routes of study at the upper-secondary and university level. They are also needed to give a future employer an indication of an applicant's level of ability. When used for these purposes, every possible measure must be taken in order to guarantee that as far as possible the marks given have the same value all over the country. The main instruments used to achieve this end are various standardized tests.

The marks given to all students in the same grade studying the same subject and, where alternatives exist, taking the same course -- "general" or "special" -- should be spread out by the mark-giving teachers according to an approximate normal distribution. It is important that this normal distribution of marks refers to the whole country. Single schools and classes usually spread differently.

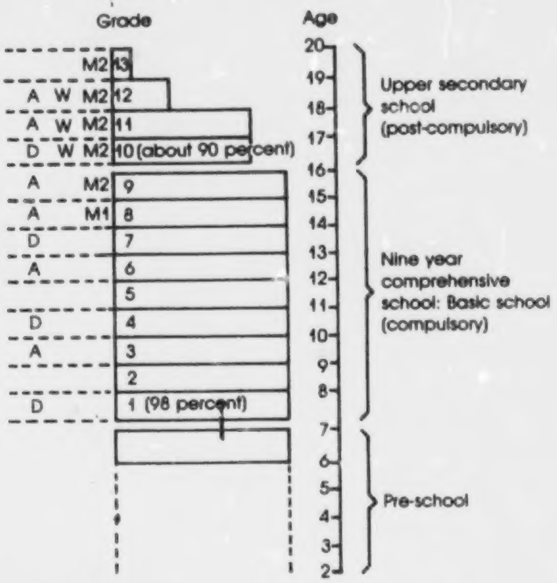
Mark	1	2	3	4	5
%	7	24	38	24	7

In the compulsory school (grades 1-9) the actual distribution follows these figures fairly well for the nation as a whole. In the post-compulsory school (upper secondary) the actual distribution of marks, due to the students choice of specialization, has gone a little "upwards", with the national means around 3.4 or 3.5.

The mark 3 denotes the mean accomplishment of the total population of students in the whole country taking the same course. Thus, the mark received by any individual student expresses to what extent he or she has succeeded in relation to that population in achieving the aims and objectives set for the subject in question. Obviously, no marking system can be perfect, in the sense that it always does absolute justice to each individual student. However, by means of the regular nationwide application of standardized achievement tests based on objective techniques, it has proved possible to go a long way towards stabilizing the marking system and eliminating variations due to change.

In the primary stage students do not get any marks. At this level local school authorities decide on other forms of information to parents and students, usually oral reports but also written non-formal reports. Marks are then given at the end of grade 8, and thereafter at the end of each term, that is, twice a year, throughout grade 9 of the basic school and the whole of the upper-secondary school. Marks given at the end of the autumn term indicate the level of

Figure 5.2. Testing and Assessment in Swedish Schools


Grade
Age
M2 13
A W M2 12
A W M2 11
D W M2 10 (about 90 percent)
A M2 9
A M1 8
D 7
A 6
5
D 4
A 3
2
D 1 (98 percent)
20
19
18
17
16
15
14
13
12
11
10
9
8
7
6
5
4
3
2
Upper secondary school (post-compulsory)
Nine year comprehensive school: Basic school (compulsory)
Pre-school

D = Diagnostic tests, voluntary
A = Standardized achievement tests, compulsory in grades 10 through 12, voluntary in grades 3, 6, 8, and 9, although used by 90% of the teachers in these grades
W = Written tests, compulsory
M1 = Marks given at the end of the school year
M2 = Marks given at the end of the autumn term and at the end of the school year

achievement reached during that term, whereas spring term marks are based on the student's performance during the whole of the academic year.

Standardized Tests

Purpose

A great number of standardized tests measuring skills and knowledge in various subjects are applied regularly all over the country. There are two kinds which are, on the whole, similar in content and construction but which differ in purpose.

One kind is used for assessing the achievement, group and individual, of the total population taking the same course. These tests are applied towards the end of a learning period. Their chief purpose is to enable the teacher to compare the performance of his own class with that of the total population and adjust his marking scale according to the outcome of the testing.

The other kind is the individual diagnostic test, used at the beginning of a teaching period, in order to give detailed profile of the student's skills and knowledge. The outcome helps teachers and students draw up a study program which will meet the specific needs of individuals and groups or of the class as a whole.

Requirements

All standardized tests have to fulfill certain requirements. They have to be valid in the sense that they actually measure the skills and knowledge defined in the aims as accurately as possible.

In principle, the tests should cover all essential aims as laid down in the curriculum. This is not possible, however, because so far no sufficiently economical and efficient technique exists for testing some aims, for example, oral proficiency.

Diagnostic tests should assess as many relevant learning objectives as possible, otherwise they fail to indicate what special measures should be taken to adjust the learning process adequately. Achievement tests can be less detailed because a nationwide reference group usually has high correlations between data obtained by measuring different abilities within the same subject. On the other hand, if an important ability is never subjected to testing, the risk is that it may also be neglected in the training program.

Achievement tests have to differentiate clearly between testees, ranking them according to their performance from top to bottom, with a high degree of reliability. What is all-important is ensuring that as far as possible the marking of these tests is uniform throughout the country, leaving no room for personal preference or bias on the part of the marker. This end is achieved either by using entirely objective techniques based on the multiple-choice principle or, where this is impossible or considered undesirable, by reducing the influence of subjective judgment to such an extent as to make it negligible.

Construction

A section of the National Board of Education[1] has until recently been responsible for the construction and distribution of all standardized tests in regular use and for instructions as to their application. Now test construction has been taken over by educational research institutes at the universities. There is for each subject a steering committee consisting of subject experts as well as experts on psychology and psychometrics. In order to ensure the necessary feedback from schools to the test makers, some committee members are active teachers. The committee is responsible for the analysis of aims and objectives necessary to secure test validity for the national school system. It is also responsible for the testing policy adopted by the schools, that is, it is responsible for the structure of the tests and for establishing principles for the choice of elements or content areas to be tested.

---

[1] The National Board of Education (NBE) is an independent body, corresponding to the administrative (non-political) branch of the Ministry of Education in most other countries.

The test-constructing institutes commission some subject experts who, as a rule, are active teachers, to construct test items along the adopted lines. The result of their work is submitted to the committee, which makes revisions deemed appropriate. The revised version is then tried out in a number of schools. About one-hundred-fifty text experts serve for short periods in temporary meetings.

The testees' answers are recorded and a detailed item analysis is made by the steering committee on the basis of data obtained by computerizing the test results. Items that have proved to be unsatisfactory as to reliability are scrapped or altered. Where computerizing is not feasible, other measures are taken to attain the highest possible degree of reliability.

In due course, the finalized version of the test battery is sent to all schools concerned, together with detailed instructions on testing procedures. The tests for the upper secondary school are compulsory. Those for the basic school are not, but about 90 percent of the teachers use them. These basic-school tests are used repeatedly over a period of some years, so they are kept confidential. In contrast, new tests are presently constructed annually for the upper secondary school; after they are used, they are published and discussed openly.

In recent years a simplified method of standardization has been practiced. This method, called "quick standardization", means that the tests are not tried out first on a representative sample of testees before they are used. The first version of the test, composed and carefully discussed by experts and steering groups, is applied directly. Replies from a representative sample of testees are then immediately collected. Norms on a five-point scale of the results are then developed by the test constructors and quickly distributed to all schools, where the teachers, who have waited for these norms for a few weeks, can now record their students' test results.

The advantages of this "quick standardization" are obvious. The try-out round can be abolished, which saves time and money. The risk of getting poor items in the instrument has proved to be minimal. A prerequisite certainly is that the test-construction experts and the steering committees are experienced test makers with a good knowledge of how different kinds of test items and instruments work at different levels of schools and different levels of student ability.

Use of Achievement Tests

The achievement tests are administered by the teachers themselves. The National Board of Education tests for the basic school and the upper secondary school in 1983 are listed in Table 5.1.

In the upper-secondary school, each test is taken in the course of one day. The dates for the different tests used in grades 10-12, which are spread over the academic year from October to April, are fixed by the National Board of Education. This arrangement has a double advantage in that the tension of final examinations crowded into a few days is avoided, and the tests occur as part of the regular school routine.

The basic school tests are administered during a four-week period fixed by the National Board of Education. Each test consists of several parts, which are taken on different days because at the lower stages many students tend to tire quickly and would fail to do themselves justice if they had to take the whole test on one occasion.

Table 5.1. Diagnostic Tests and Achievement Tests in Sweden

A = Achievement Test, average time allowance, 180 minutes
D = Diagnostic Test, average time allowance, 80 minutes

Basic School

Subject	Grade					
	3	4	6	7	8	9
Swedish	A	D	A	D		A
Mathematics	A	D	A	D		A
English		D	A		A	
French/German						A

Upper-secondary School

Subject	Three-to-four-year streams			Two-year streams	
	Grade				
	10	11	12	10	11
Swedish			A		A
Mathematics	D		A	D	
English		A			A
French/German		A			
Chemistry	D	A			
Physics	D	A	A		
Mechanics				D	

Table 5.2. Compulsory Written Tests in Swedish Upper-secondary Schools: Three- and Four-Year Streams, Recommended by the National Board of Education

Subject	Grade		
	10	11	12
	Number of Tests*		
Swedish	2	4	3
English	4	3	
Commercial English			2
French/German	2	3	4
Commercial French/German			2
Mathematics	4	5	3
Technology	1	2	
Physics		3	3
Economics		3	2
Accountancy			3
Machine/Building Construction			2
Building Technology			2
Electricity			4

*The time allowance is 90-180 minutes for test.

The teacher records each student's test result in the form of a mark on the five-point scale and works out the average mark for the whole class. Towards the end of the term, when the final marks are to be set, these data are used to see if the teacher has a "good" class or a "poor" class or if the class is homogeneous or heterogeneous in achievement.

Non-Standardized Tests

The curriculum emphasizes that the main purpose of compulsory written work is to provide opportunities for practice and that, when used as a basis for marking, students' results on written tests must not outweigh other important factors. The NBE has recommended that compulsory written tests be given for three-and-four-year streams in upper secondary school. Each test takes 90 to 180 minutes. The NBE recommends that teachers in upper-secondary schools give one to four such written tests a year. A reason to give these tests is to make sure that not only the central curriculum but also local and individual areas become subjects of assessment.

The same principle applies to non-compulsory written exercises, which are used at all stages throughout the school system. It is up to the teacher to decide on the nature and quantity of such exercises, which are normally an integral part of ordinary classroom work.

Sometimes, however, all the students in a school taking the same course take the same written test simultaneously, and occasionally this arrangement is extended to include all the schools in a district. In such cases a team of teachers work out marking norms, which are then used by all the teachers involved. This enables them to compare the standard of their own classes with that of the others.

The Assessment Process

Classroom Observations

The teacher's main task is, of course, to aid the students in their personal development and to help them acquire the skills and knowledge defined in the aims laid down in the curriculum.

This entails continually assessing the students' work and keeping them informed of their progress. Teachers are therefore advised to observe each individual's performance within the class and to record their observations from time to time.

All performances must be taken into account, and the teacher must be on his guard against paying too much attention to results that are easier to assess than others. It is particularly important to take proper account of oral proficiency, in the mother tongue as well as in foreign languages, since this most important ability cannot at present be easily measured by means of objective techniques.

The upper-secondary school class used to be visited occasionally by a subject expert. These experts study the work in progress and discuss it with heads, teachers and students, both in conference and privately. They are thus able to form a good overall picture of all school activities concerning their subjects and of the general standard of skills and knowledge achieved in different schools, as well as to give advice on teaching methods and evaluation. In the basic school the same functions are performed by other categories of inspectors and advisers.

Written Tests

The teacher keeps a record of each student's performance in all written tests taken during the evaluation period. In the upper-secondary school all compulsory test papers are filed so as to be available for principals and visiting inspectors. By examining the papers, they are able to see if the marking principles applied by the teacher tend to be more lenient or severe than the average and are thus in a position to assist teachers in their endeavor to attain a high degree of uniformity in assessing the student's work.

Final Assessment

Towards the end of the term, the teacher surveys all the evaluation data collected and ranks the students from top to bottom according to their individual level of ability, giving each a mark on the five-point scale. These marks are preliminary and may have to be adjusted. As stated above, the curriculum emphasizes that the student's standard of performance within the class must be given proper weight in relation to their results on written work. In the job of assessing the students' overall standard, the teachers will find their task greatly facilitated if they have kept a running record of their classroom observations.

The main function of the standardized test is to be instrumental in achieving the highest possible degree of uniformity in the marking system. A detailed description of the procedure to be followed is contained in the curriculum.

First, the teacher calculates the mean of the preliminary marks and records their distribution over the five-point scale. Then he compares these data with the mean and distribution of marks obtained by the class in taking the nationally standardized test. If the two means are identical, or if the difference between them does not exceed ±0.2 (which was seen as an acceptable tolerance for chance influences), the teacher can conclude that the preliminary marks indicate the standard of the class correctly in relation to that of the total population. If the two distributions also coincide more or less completely, the preliminary marks can be taken as final.

Each teacher delivers the marking documents to the headmaster's/headmistress's office; all the relevant data are arranged and recorded in such a way as to facilitate comparisons between classes and within each class. This material is available at a meeting, called a class conference, which is attended by the head and all the teachers taking the class in question for one or more subjects. The purpose of the class conference is to make final decisions on the means and distributions of marks. Comparisons are made between the standard achieved in different subjects and also between the achievements of different classes in the same subject. A teacher who wants to retain noticeable differences between test results and preliminary marks has to convince the class conference that there is a valid reason for doing so.

The adjusted means and distributions of marks for those subjects in which standardized tests are taken are used as guidelines for adjusting the means and distributions for other subjects. This principle is based on the well-known fact that within a class the means and distributions, as a rule, have a fairly high degree of correlation, regardless of subject.

Dividing up the marking procedure into two steps, one for preliminary marks and a later one for final marks, is important. The class conference between these two steps aims at making single marks for single students comparable all over the country. This way it has become possible to base the selection for higher studies on secondary-school marks instead of university entrance examinations.

Free Choice of Study Route in School

In grades 1-6 (primary stage) all pupils follow the same courses. In grades 7-9 (lower secondary) 85 percent of the schedule and the courses are the same for all. During the remaining 15 percent of the time students can choose elective subjects. Two-thirds of them usually take a second foreign language. The rest take extra courses in other school subjects. The choice is free for the students, regardless of their results. At the end of grade 9 the students get a final grade, saying that they have completed their compulsory education and listing the subjects they have studied and their results on the five-point scale in each of these subjects.

About 90 percent of the students from compulsory school transfer to the voluntary upper-secondary school. This school can take all students (the number of places is actually 115 percent of the number of students leaving compulsory school. Even here the students in principle have free choice. There are twenty-two specializations to choose among, one technical for four years, four theoretical for three years, three general for two years, and fourteen vocational for two years. Usually 80-85 percent of the students take one of these twenty-two specializations. The remaining 5-10 percent of the applicants take so-called special courses, which have a length from half a year to three full years. Only two out of three beginners can get the study specialization or special course they have named as their first choice. The others have to accept a study route that they indicated as their second or later choice. The reason for this is that certain routes are popular and have more applicants than places, whereas other routes have more places than firsthand choices.

Where a selection is made, the mean mark from the final grade of the compulsory school makes up the rank order. This mean mark includes all subjects, including art, music, and physical education, as well as academic subjects. Here something unforeseen has happened in recent years. The most attractive study specializations are no longer the traditional pre-university specialties (science, technology, humanities, social science, and economics) but certain vocational ones. It is now in fact more difficult to get a place in nursing, auto mechanics, or the agrarian or electro-technical specialties.

Selection for Higher Education

In 1975 the Swedish Parliament decided that the eligibility requirements for higher education should be broadened to include not only the traditionally three-and four-year academic lines of upper-secondary school but also students from the two-year specializations of this school (including vocational training) as well as adults in gainful employment. According to this decision new selection rules were introduced in 1977, which means the allocation of a quota proportional to the different applicant categories and changed criteria for selection between eligible applicants.

Broadened eligibility means that persons who have completed at least two years of upper secondary school studies or have equivalent qualifications (for example, folk high school[2] studies) are generally eligible for higher education as are all persons over the age of twenty-five with at least four years' work experience (the so-called 25:4 rule). The general eligibility requirements include a knowledge of English and Swedish equivalent to two years of upper-secondary studies (students eligible on the basis of age and work experience are automatically assumed to fulfill the requirement for a knowledge of Swedish). However, it is important to note that these are only the general eligibility requirements, which must be fulfilled by all applicants irrespective of the study program. In addition, specific upper-secondary-level knowledge of particular subjects is generally required, the so-called special eligibility requirements, which vary from course to course. In order to be admitted to medical training, for example, an upper-secondary-level knowledge of several natural sciences is required. Adults without a complete upper-secondary education can acquire these subjects through municipal adult education, which is widespread in Sweden.

Where competition for places exists, however, some form of selection must occur. The rules of selection differ from one type of program to another. For instance, for single independent courses (in language, economics, computer theory, or business administration), many students already have completed vocational training or have an academic degree. Admission is made on the basis of student interest in or need of the course. In the case of teacher training, students are admitted on the basis of "qualification points" rather than their needs or interests. In the case of full degree programs, applicants are divided into four groups according to eligibility:

(1) Applicants who have completed three- or four-year programs of upper-secondary education.
(2) Applicants who have completed two-year programs of upper-secondary education.
(3) Applicants with an equivalent education from a folk high school (two or three years).
(4) Other applicants, particularly those eligible on the basis of age and work experience.

A maximum of 10 percent of the places are reserved for foreign students or for admittance on special grounds (usually called "quota group five").

Each group is allocated a number of places corresponding to the number of applicants in the group, the so-called proportional quota allocation. The applicants in each group then compete among themselves.

---

[2] The folk high school is a type of adult training institute typical in Scandinavian countries during the last hundred years. Most of them are run by so-called popular movements (labor movements, religious movements, temperance movements, and sport movements, for example). Sweden now has 120 such schools with approximately 15,000 full-time students in two- or three-year courses plus approximately 200,000 students in short courses (from weekend courses to eight-week courses). They are boarding schools. They were earlier a substitute for secondary education used by adults in farming and industry. Now they mostly make up an alternative to regular secondary schools or specialized post-secondary training.

In quota groups one to three, selection takes place on the basis of the mean mark from upper secondary school on the five-point scale, plus additional points for work experience. In quota group four, selection is made on the basis of the results of a study aptitude test and work experience. The study aptitude test is of the same type as the American Scholastic Aptitude Test and other corresponding university and college entrance tests, including also subtests for understanding English. School marks yield a maximum 5.0 points, and work experience, 2.0 points. Even the study aptitude test in quota group four yields a maximum of 2.0 points.

Table 5.3. shows the distribution in percentages of admitted students to full-degree programs with restricted entry in 1977. Since this year, when the quota system was fully implemented for all Sweden, the proportions of the quota group has not changed very much.

Table 5.3. Distribution of Students in Full-degree Programs, 1977

Quota Groups	University Students in General	Medicine	Technology	Teacher Training
(1) Three-year secondary school (plus work experience)	73	66	88	66
(2) Two-year secondary school (plus work experience)	10	8	2	19
(3) Folk high school (plus work experience)	2	1	2	5
(4A) Adult students (25:4) without school marks	4	9	2	4
(4B) Adult students (25:4) with school marks	3	9	2	3
(5) Foreign students (and students admitted on special grounds)	8	9	8	3
Total Percent	100	100	100	100
Number of Students	10,777	1,390	4,091	487

One circumstance which complicates the procedure is that it is possible to be eligible in several groups. The most common form of double eligibility is that the applicant is eligible in quota group four and at the same time has qualifying school marks. Only a small proportion of quota group four is eligible purely on the basis of age and work experience and has no upper secondary school marks. This is why, in table 5.3., quota group four is split up into two subgroups, 4A with "pure" 25:4 applicants and 4B with 25:4 applicants with double eligibility.

These figures tell us, that the "older students" (twenty-five years or more, with a minimum of four years of work experience) are relatively few (7 percent) and that some of them (3 percent) could have applied according to their school marks. The majority of "older students" are not found in these full-degree university programs but in the so-called single courses of one year or shorter. There the "older students" usually make up more than half the number, especially in extension courses and extramural courses.

It is worth noting that no fewer than one third of the new students usually has previously been enrolled in higher education. During all the 1970s the percentage of mature students and students with non-traditional backgrounds in higher education in Sweden has gradually increased but is still a small number. A certain redistribution has occurred between various groups of students and between different types of study programs within higher education. These are primarily the effects of the new selection rules and only secondarily the result of opening up higher education to new groups.

The fact that young students coming directly from upper-secondary school now have to compete for admission with older applicants who not only have extra points for work experience but sometimes also have a full academic degree, is matter of controversy. In 1980 Parliament decided that a minimum of one third of all available places in full-degree programs should be reserved for applicants coming directly from upper-secondary school. In fact, about two thirds of the places are given to these applicants.

Sweden seems to be one of the very few countries where the majority of students entering universities and other types of higher education are selected using marks from secondary school, *not* a special entrance examination test. In attempting to improve student selection, the main stress has been on making school marks comparable across the country. For this the use of national standardized achievement tests in schools and the introduction of the two-step procedure for setting preliminary and final marks are the two cornerstones. It is also worth repeating here what was said at the beginning, that is, that Sweden is a small country with a linguistically, culturally, and socially fairly homogeneous population and a recently unified school organization, circumstances which have facilitated this method of selection for higher education.

Individual Results as Indicator of School Results

Assessments of outcomes of education in Sweden have so far focused mainly on results of individuals in single grades, subjects, or courses. Test results of this kind tell whether the individual has passed or failed, what grade he has achieved, and how he stands in relation to other individuals in the population concerned.

This kind of individual assessment and evaluation has existed since 1944 and is fairly well developed. It is important to note, however, that this kind of evaluation of individual results is concentrated on what are called "skill subjects" or "instrumental subjects." These are given the compulsory-school level in Swedish (reading and writing), mathematics, English, German, and French (the two last languages as elective subjects). On the post-compulsory, upper-secondary level, individual tests are given in the same subjects as well as physics and chemistry.

No centrally developed and administered tests are given for so-called "orientation subjects" (different courses of social studies and science) and vocational subjects. The reason for this is that these subjects and courses usually do not have a uniform national structure.

A number of research and development activities have begun to extend this kind of individual evaluation to evaluations of whole educational programs. An educational program concerns many students, subjects, and courses. A program can be training for a trade, profession, or a restricted examination. Such programs usually also have specialized evaluation criteria to be applied not only to individual students but also to students as a group. An educational system is an even wider concept, that is, primary school for the whole nation or for a region.

In the post-war period a number of other steps were also taken to improve education and adapt it to new individual and societal demands. For instance, through extending the period of compulsory training at school, broadening secondary education to include nearly everybody in each year cohort (from elite education to mass education), coordinating general-academic and technical-vocational education into new programs, and making the school structure comprehensive with different programs within the same organization. The organizational diversification of students into different fields of study was postponed, final examinations were changed, and access to higher education became easier.

These changes (the list could be extended and specified) articulate the need for an evaluation which goes beyond an assessment of individuals in single subjects. Two types of more complex evaluation are now being tried out. One type can be called a "total curriculum" evaluation. The knowledge, skills, and attitudes of the individual student, seen as a unit, make up the subject of the evaluation. A second type is the result of a group of students in one or more evaluation criteria. In this second kind of evaluation, differences between individuals, schools, and regions become important subjects of analysis.

Program evaluation and system evaluation used to be done predominantly through inspection, accreditation, and specially designed reviews by evaluation institutions. The ambition is now to develop national assessment programs, in which the data from individual tests will make up the statistical basis.

Two types of scores of assessments have proved to be of utmost importance: (1) the mean score of the institution and the system in whatever criterion or type of assessment is defined; and (2) the variable (the standard deviation or other kinds of corresponding estimation) of the same assessment. The first one indicates the level of the institution or the system. The second one tells how many students fall below an acceptable minimum standard (and therefore might need extra support) and how many are especially high (and therefore perhaps should be observed and treated in a special way). Quite naturally it is important to have not only traditional achievement test scores but also estimations of different kinds of non-cognitive results of education.

Considerations of this kind, emerging from egalitarian goals, underline the need for different analyses of variance, such as: inter-individual, intra-institutional (classes, groups), inter-institutional, (schools), and intra-system variance (regions, districts).

This is supposed to help teachers, heads of schools, administrators, and policymakers find out the strong and weak points of a system or an institution and how available resources might be used to improve the results.

Such analyses of variance are easily done with assessments according to standardized achievement tests. In principle they also could and should be done with criteria of a non-cognitive kind, for example, the kind and degree of student participation in decision-making within institutions and systems and how training within institutions and systems is related to changes in society, industry, and the labor market. Obviously, "effectiveness" can also be defined in many, often quite pragmatic ways, as for instance the pass-fail frequency, the frequency of class-repeating, the annual cost per student, and the total public expenditure per graduated student.

6

A BRIEF INTRODUCTION TO THE SYSTEM OF HIGHER SCHOOL ENROLLMENT EXAMINATIONS IN CHINA

Lu Zhen

Historical Background

The Recommendation System in Ancient China

For fourteen hundred years, from 840 B.C. to 587 A.D., selection and promotion to positions of authority in many ancient Chinese dynasties were based on a system of recommendations by educational institutions at various levels.

As far back as 840 B.C., when the West Zhou Dynasty had just been founded, the ruling classes, pressed by administrative needs to consolidate power, established the gongshi system and the system of selection and promotion from the level of xiangxue (schools on the county level) through guoxue (schools on the prefecture or province level) to the level of taixue (the level of national university). The enrollment examinations included moral conduct and military feats. The enrollees were predominantly princes and the eldest sons of warlords and senior officials with the exception of a few rare talented plebeians.

With the crumbling of the slave system and the establishment of various feudal kingdoms, the seventh century B.C. (that is, the Spring and Autumn Period) witnessed the emergence of the system of retaining literati. To expand their sphere of influence, these feudal kingdoms endeavored to attract literati and encouraged the running of private schools. Some kings and private school teachers had thousands of students. For instance, Confucius, the famous educationist of the Lu Kingdom, was reputed to have three thousand students, only seventy-two, or 2.4 percent, of whom were considered to be his "worthy students."

A system of recommendation was adopted from 344 B.C. to 605 A.D., that is, from the Reform Movement of Shangyang in the Qin Kingdom and the unification of China by the First Emperor of the Qin Dynasty to the beginning of the Sui Dynasty. Under this system, each prefecture kingdom (which was comparable to a province) recommended a certain number of people in accordance with certain regulations. They would then be recommended by the prime minister or other high officials as xiaolian (a man of filial piety and honesty) or xiucai (a literary talent). After certain examinations of verification, they would get their appointment as officials.

These systems, despite their differences in name and the requirements of the candidates, share a common trait: that is, the method of recommendation. Historically, these systems played certain positive roles, as they promoted a number of politicians and reformists. However, they also entailed numerous social evils, mainly the monopoly of recommendations by the bureaucrat-landlord class. Generally, recommendations were not made according to the candidate's virtues and talents but according to his family status. Consequently, many capable people were deprived of the opportunity to go to school because they were children of plebeians. Sometimes it even happened that a person was recommended to be a xiucai without knowing a single word. Very often, no high officials were from poor origins, and even low officials had aristocratic genealogies. In short it was a corrupted system. On account of its flaws, the system of recommendation was abolished by the Sui Dynasty in 587 A.D.

The Imperial Examinations System in Ancient China

The year 606 saw the establishment of the Imperial Examinations System, which continued for more than thirteen hundred years until 1905, the end of the Qing Dynasty. "Imperial Examinations" in China involved selection and promotion according to examinations of different levels. The exact methods varied from dynasty to dynasty. In the Tang Dynasty, examinations were held at two levels, one under the charge of prefecture authorities and the other administered by the Ministry of Rites (which includes the present-day Ministry of Education). Those who passed the examinations were appointed as officials. In the Sung Dynasty, a final imperial examination was added, which the emperor himself presided over. In the Ming and Qing Dynasties, four levels of examinations were given, namely, (1) the pre-prefecture examination, the passers of which were given the title xiucai; (2) the prefecture examination, the passers of which were called juren; (3) the examination held by the Ministry of Rites, the passers of which were given the title gongshi; and (4) the final imperial examination, with the emperor himself as the presider, the passers of which held the title jinshi. Among them the number one scholar was called zhuangyuan, the number two scholar was called bangyan, and the number three, tanhua; all were appointed official positions.

The imperial examination consisted of five subjects: (1) jinshi (candidate in the highest imperial examinations) prose, poetry, and political essays, mainly testing the level of literary writing and political argumentation; (2) classics, mainly the Five Classics (namely, the Book of Songs, the Book of Rites, the Book of Changes, and the Spring and Autumns Annals), mainly testing the candidates' comprehension of the essence of those classics; (3) law, testing the candidates' knowledge of law; (4) calligraphy; and (5) mathematics.

The examination questions generally took the following forms:

(1) tiejing, namely, filling in the blanks. The text of the classics was covered, leaving exposed only some characters in a certain line. The examinees were required to fill in the rest of the words;

(2) answering questions concerning the classics, in both oral and written forms;

(3) writing poems on assigned topics;

(4) discoursing on politics, that is, writing political essays;

(5) jingyi, writing an essay expounding on profound truths in politics, economics, morality, and self-cultivation by applying the general principles of the classics, after being given a particular sentence in one of the classics. Starting from the Sung Dynasty, more and more emphasis was laid on this. In the Ming and Qing Dynasties, the essays were limited to "The Four Books" and "The Five Classics", which gradually resulted in the notorious sterotyped writing known as the "eight-part essay".

In comparison with the previous systems of recommendation, the imperial examination system was a great step forward. By dint of the different levels of examinations, which had some fairly objective standards, this system made it difficult for fraudulent practices and thereby promoted many extraordinary politicians and men-of-letters. However, due to the corruption of the social system, both the form and content of the examinations gradually lost their flexibility. Finally, when they came to be dominated by "the eight-part essay", which required only memorization, the examinations lost their usefulness in selecting and promoting capable people.

The Higher-school Enrollment Examination Before the Revolution

After the Opium War in 1840, the tentacle of western culture found its way into China. Some progressive intellectuals then started criticizing the imperial examination system as "empty and useless". Meanwhile, they called for the abolition of "the eight-part essay" and the establishment of new schools.

From 1895 to 1898, famous schools, namely, Zhong-xi School, Nanyang School, and Jingshi University were built in Tianjin, Shanghai, and Beijing, respectively. In 1912, that is, right after the Revolution of 1911 which overthrew the Qing Dynasty, the Ministry of Education renamed these schools as universities.

In 1917 the Ministry of Education stipulated that all students of normal universities and colleges must pass an enrollment examination. In 1938 the Ministry of Education announced that from then on unified enrollments would be held in all normal universities and colleges and that such matters as the subjects of examination, the assignment of questions, and the standards of acceptance were to be under the charge of a unified enrollment committee. According to the announcement, both written and oral examinations would be given. The written examinations were divided into three groups for different examinees. Each group consisted of seven subjects, among which civics, Mandarin Chinese, English, and the history and geography of China were common to all the groups. Mathematics was also an obligatory subject, but the examination paper varied slightly from group to group. In addition, examinees of the first group were required to take history of the world and choose one from the three subjects, that is, physics, chemistry, and biology; the second group was required to take physics and chemistry; and the third group was required to take biology (or world geography for candidates of the department of geography) and choose between physics and chemistry. The oral examination was put in the charge of relevant schools. Unfortunately, this system of unified enrollment examination was never put into practice because of the War of Resistance Against Japan. In 1947 China had 125 higher schools with only 150,000 students. All these schools had their own systems of enrollment examinations.

The Higher-school Enrollment Examinations After Liberation

When the People's Republic of China was founded in 1949, there were altogether 205 higher schools with 117,000 students. In 1952 some adjustments were made, and the sixty-five private higher schools were turned into public schools.

To meet the needs of the planned economic construction and fulfill the plan of higher school enrollment, a national higher school enrollment committee was set up in 1952. This brought about the unification of higher school enrollment throughout the country (with the exception of Taiwan), including the unification of examination questions, requirements for application, examination subjects, general principles, policies, and concrete measures of enrollment. A working committee of higher school enrollment was subsequently established in each of the provinces, cities, and autonomous regions (except for Taiwan). The committees were in charge of applications, examinations, political examination, physical examination, grading examination papers, and admissions. This work was done according to regulations stipulated by the national committee. This guaranteed the fulfillment of the country's higher school enrollment plan and the quality of the enrollees. It also brought convenience to the examinees, who could then take examinations for any higher school in the country without having to travel beyond their counties, thus saving considerable money and manpower.

Examination Subjects

Since liberation, some changes in the subjects of examination have taken place. In 1952 there were eight subjects: the ABC of politics, Mandarin Chinese, a foreign language (either Russian or English), history and geography, mathematics, physics, chemistry, and biology. In addition, applicants to departments of physical education, music, and painting were examined in the relevant subject.

In 1954 the enrollment examinations were classified into two categories in order to suit the needs of different specialities. First, applicants to science and engineering, medicine, agriculture, and forestry departments were required to take the following seven examinations: the ABC of politics, Mandarin Chinese, mathematics, physics, chemistry, biology, and a foreign language. Second, applicants of liberal arts, law, finance and economics, and physical education departments were required to take five examinations: the ABC of politics, Mandarin Chinese, history, geography, and a foreign language. Applicants to the finance and economics department were additionally examined in mathematics, and the applicants to the department of music, painting, and opera were also examined in the relevant subject.

In 1955 the examinations were further classified into three sets. Applicants to science and engineering departments were examined in five subjects: Mandarin Chinese, the ABC of politics, mathematics, physics, and chemistry. Applicants to the departments of medicine, agriculture, forestry, biology, psychology, and physical education were also examined in five subjects: Mandarin Chinese, the ABC of politics, the fundamentals of Darwinism, chemistry, and physics. Applicants to the departments of liberal arts, law, and finance and economics were examined in four subjects: Mandarin Chinese, the ABC of politics, history, and geography. In 1958 a foreign language was added to every category.

In 1964 the subjects of examination again fell into two classes. Candidates of departments of natural science (science and engineering, agriculture, forestry, and medicine) were examined in Chinese, the ABC of politics, physics, chemistry, and a foreign language. Candidates of the departments of liberal arts were examined in Chinese, the ABC of politics, history, geography, and a foreign language. In addition, would- be finance and economics majors were examined in mathematics, and foreign language majors also took an oral examination.

Due to the Cultural Revolution, the unified national higher school enrollment system was suspended from 1966 through 1976. From 1966 through 1971 higher school enrollment was totally stopped. When higher school enrollment resumed in 1972, middle school graduates were excluded. Only workers, peasants, and soldiers with at least two years' practical experience and a junior-middle-school graduate level were qualified for enrollment. All academic examinations were cancelled, and a new system of recommendation was adopted. Consequently, fraudulence became a common occurrence, and the quality of fresh enrollees and the level of education drastically deteriorated. Our cause of education and talented training suffered a serious setback.

In 1977 the recommendation system was nullified, and the system of unified national higher enrollment was again called into operation. Fresh middle-school graduates again became the main source of enrollment, and the system of academic examinations was resumed. We adopted the principle of "admission according to quality" with regard to the candidates' moral character, academic level, and health, thereby greatly improved the quality of the enrollees, restored the normal order of teaching, and ameliorated the general mood of society.

Since 1978 the unified national higher school examinations have been classified into two groups:

(1) Candidates for the liberal arts major take examinations in six subjects: political, Chinese, mathematics, history, geography, and a foreign language;

(2) Candidates for natural science (science and engineering, agriculture, forestry, and medicine) majors take examinations in seven subjects: politics, Chinese, mathematics, physics, chemistry, a foreign language, and biology (which was added in 1981).

As for the foreign language, the examinee can choose from English, Russian, Japanese, French, German, and Spanish.

Every year the unified examinations take place on July 7, 8, and 9. The time limit for each examination is normally 120 minutes, with the exception of Chinese (150 minutes) and biology (60 minutes). The time of examination for all the subjects is stipulated as follows:

July 7:	Morning:	Chinese	Afternoon:	Chemistry/Geography
July 8:	Morning:	Mathematics	Afternoon:	Politics/Biology
July 9:	Morning:	Physics/History	Afternoon:	Foreign Language

According to the regulations, all the examination places are located in cities of county level or above. The organization and leadership of the examinations, the selection of the examination places, the regulations for proctoring the examinations, are all strictly stipulated.

The Goals of the Examination

The main goal of the examination is to select qualified enrollees for higher schools. In recent years more than two million new middle-school graduates and other young people have taken the examinations. This year only about 560,000 students, or about one-fourth of the examinees, were enrolled. Therefore, the examinations are highly competitive. In view of the fact that most of the examinees graduated from middle-school some time ago, the examination papers must not unduly favor them over the actual level of the average current middle-school graduate.

We stipulate the following principle in setting examination questions: they must be both advantageous to the selection of enrollees for the higher schools and conducive to the improvement of the level of teaching in the middle schools.

In accordance with the above principle, there are four points in setting examination questions. First, the higher-school enrollment examination, being a kind of selection examination for universities and colleges, must have a high predictive validity. The higher-school enrollment examination is not an achievement examination for middle-school graduates; rather, its main purpose is to select only part (about one fourth) of a great number of middle-school graduates for the pursuit of advanced studies in higher schools. Therefore, a considerable section of the examination questions must be tolerably more difficult than those of senior-middle-school leaving examinations.

As regards the contents, the emphasis is on testing the basic knowledge and skills which the examinees have learned in middle school and their ability in applying them in a living way. In short, both knowledge and ability must be taken into consideration in setting the examination questions. To take the examination of Chinese for an example, the examination papers are designed to test the examinees' reading comprehension, writing, and essential knowledge about the language by questions about present-day Chinese and composition. Normally, reading comprehension and writing constitute fifty points each, and basic knowledge, accounts for twenty points. To take another example, questions of mathematics, physics, and chemistry, which fall in the scope of mathematical knowledge, are mainly designed to test the examinees' basic knowledge and skill in these subjects, their ability to analyze, to reason, to calculate correctly and quickly, and to apply their knowledge in a comprehensive way.

Secondly, as the higher-school enrollment examination has some influence on teaching in the middle schools, it is essential that the scope of the examination questions do not go beyond that of the syllabuses of the subjects for middle-school students, nor exceed that of middle-school textbooks. Furthermore, they should cover as much as possible the basic contents of the middle-school syllabus and textbooks as well as their requirements for basic skills. For example, in the chemistry examination paper of 1984, 17 percent of the questions came from the third-year junior middle-school textbooks; 40 percent came from the first volume of senior middle-school texts, and 42 percent from the second volume. The paper included fifty concepts and over eighty elemental compounds, which covered all the basic theories in senior middle-school textbooks, including physical constructions, the periodic laws of elements, chemical balance, as well as chemical calculations.

The third point in setting examination questions is that it is also important that difficulty in the examination questions be suitably graduated. Judging from the higher-school enrollment examinations in the past few years, it is essential to give some questions of intermediate levels, in order to distinguish the one-fourth expected enrollees from the rest. Even these questions should be of varying degrees of difficulty so as to bring out the difference in the high scorers' grasp of knowledge and their ability to put it into practice and to widen the gaps between their scores. This will enable us to pick out the qualified examinees according to their academic merits. In the meantime, due consideration must be given to the actual conditions of teaching in the middle schools to guarantee a proper rate of passers. Therefore, the questions should be of various levels and arranged from the easiest to the most difficult.

Lastly, in view of the fact that middle schools in China are divided into key schools and ordinary schools with their respective requirements of education, in 1984 we decided to attach some optional questions of greater difficulty and a higher level than the requirements of the middle-school syllabuses to the examination papers of mathematics (for students of natural science), physics, chemistry, English, and Russian. The optional questions for each subject added up to ten points, but they were not added to the total score. They are used only for the reference of higher schools in admitting new enrollees.

In recent years, the form of the examination questions has constantly been improved. The level and scientific method have also been raised. In respect to the types of questions, new experiments have been made every year in addition to such traditional types as true-or-false questions, filling-in-the-blanks, multiple choice, filling-in-the-pictures, reading-the-pictures, calculations, defining terms, correcting mistakes, answering simple questions, and comprehensive experiments. Each of these plays some role in testing the examinees' level of knowledge as well as their ability to apply their knowledge in a living way and to make judgments and inferences. At present China is paying more and more attention to the standardization of examination questions, the objectivity of grading, and the modernization of the whole testing system in higher-school enrollment examinations. For instance, in recent years, multiple choice, which is still under special experiment, is being introduced into higher-school enrollment examinations, and its merit of one correct answer and objectivity of grading is being assimilated into other comprehensive questions as an experiment. Nevertheless, just as the traditional questions and answers have their own limitations, it is difficult to use multiple choice alone to achieve the same effect as that produced by the classics and the "political essay."

In setting the examination papers, we always prepare two sets of questions of roughly the same level for each subject, that is, Set A and Set B, together with their respective correct answers. Set B is prepared in case the examination is postponed on account of some unsurmountable difficulties.

Grading the Examination

Printing and distributing the higher-school enrollment examination papers, and the examination itself, are under the unified management of the working committees of higher-school enrollment of all the provinces, municipalities, and autonomous regions (except for Taiwan). There are altogether approximately 53,000 examination places in some 2,400 counties, cities, and districts throughout the country. On July 7, 8, and 9, more than 100,000 people are needed to proctor, and after the examinations another 100,000 teachers must be organized for grading the examination papers, which normally takes about two weeks.

In order to strengthen the management of grading and keep unified standards of grading, all the grading workshops are located in a small number of higher schools in each province or city. Normally, examination papers of the subject are sent to one school.

The grading work is mainly done by higher-school teachers with the participation of a proper percentage of experienced middle-school teachers. Normally, university teachers account for 70 percent of the examiners, and middle-school teachers account for 30 percent. Before the grading work starts, all the examiners are organized to study the "keys to the questions" and the "standards of grading." On the basis of this, some trial grading is done to train the teachers. In the process of grading, every examination location has some senior teachers to do the checking work. Meanwhile, people designated by the enrollment committees of the provinces, municipalities, and autonomous regions selectively examine the graded papers in order to eliminate all possible mistakes and unfairness and to ensure the quality of the grading.

The grading is done in a flow process, each examiner being in charge of one question (composition and essay question are under the common responsibility of two examiners). The expenses of the grading are covered by the national enrollment expenditure. The examination scores are totally confidential, accessible only to the examinees themselves.

Admissions

The decision to accept new students is made under the unified organization of the enrollment committee of the provinces, municipalities, and autonomous regions. After the political and physical examinations, we make a decision on the basis of minimum scores. This identifies about 10 percent more than the total number of the students to be enrolled according to the plan. The universities and colleges will then select their enrollees according to their scores in relevant subjects and the order of preference in the students' applications (each student can apply for two schools, plus three more for reference; for each, he can apply for two departments or specialities.

To make up for the shortcomings of the unified examinations, we have decided that starting this year, a small number of students with excellent academic records or of rare intelligence be exempt from the examinations and sent directly to the universities or colleges and that a few other students with specialized skills or knowledge will be examined on basis of recommendations.

From this brief survey of the system of selection and promotion in the history of China, it is clear that as a main means of enrollment, the examination is far more reliable and fair and much less susceptible to social evils. The system of the unified higher-school enrollment examination is very effective in selecting qualified middle-school graduates and other young people for universities and colleges, in accordance with the overall enrollment plan of the state. It ensures the quality of the new enrollees of the higher schools and the training of talented people and, to certain extent, it also gives an impetus to the improvement of teaching in the middle schools. In the past few years, we have made some improvements in the subject, content, and form of the examinations and in the scientific quality of setting the questions. However, the present traditional examinations still have a long way to go to meet the demand of standardization and scientific methods, and they do not quite suit the needs of our "four modernization" program. Therefore, we shall make full use of the materials submitted at this seminar and learn from other countries' strong points on the basis of our own experience, so as to improve our higher-school enrollment examinations and strive to establish a higher-school enrollment examination suitable to our national condition in the near future.

7

DESIGNING THE ENGLISH LANGUAGE PROFICIENCY TEST IN CHINA

Gui Shichun

Introduction

In a large country like China, standardization often involves the use of immense resources of manpower and financial support. This is the situation facing the administration of standardized tests in China. The topic that I choose for discussion is that the ideas and procedures of a standardized test can be tried out with the help of a microcomputer. In this way we can follow all the technicalities of standardized testing, which have become so specialized that only highly professional testing organizations can handle them properly. I believe that only when more and more people are familiar with the concepts of standardized testing and the sequential steps of constructing a test in a small way can China have standardized tests in a big way. Over the past few years, I had the privilege of working with a team of test-setters on the English Proficiency Test (EPT), the first of its kind that had been trying to follow all the requirements of a standardized test. I would like to make use of the data we collected to illustrate my points. I am addressing this to Chinese readers with the intent of familiarizing them with the ideas and techniques of modern testing. In the course of my preparation, I have made extensive use of such classics as Robert Thorndike's *Educational Measurements*.

Stage One: Planning a Standardized Test

Constructing a Test

Figure 7.1. illustrates the stages in constructing a test.

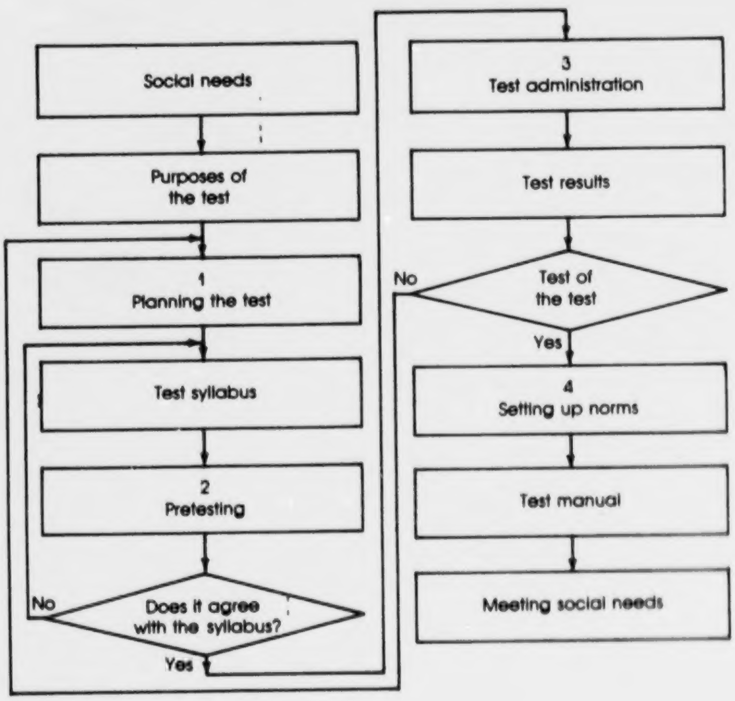
Choosing the Right Test

We can see from Figure 7.1. that stage one is a very important step in the sense that it is fundamentally a decision-making process. The test planner is faced with the question of studying the needs of the society in terms of educational objectives and deciding on the kind of suitable test to fulfill such needs. In general, tests can be categorized in different manners: according to their use, the way test scores are reflected, the manner in which the items are graded, the way the skills and abilities of the students are measured, and the test requirements.

1. According to their use

An achievement test measures the progress or achievement the students made over a period of time. It has to follow rather rigidly the course syllabus,

Figure 7.1. Flow Chart of Stages of Standardized Test


Social needs
Purposes of the test
1 Planning the test
Test syllabus
2 Pretesting
Does it agree with the syllabus?
No
Yes
3 Test administration
Test results
Test of the test
No
Yes
4 Setting up norms
Test manual
Meeting social needs

and the scores of the students reflect what they have already learned of that specific syllabus.

A proficiency test measures the students' proficiency level in particular skills or areas of knowledge so as to determine whether their level corresponds to specific requirements. In the case of a language proficiency test, adequate control over language skills is implied. The proficiency test differs from the achievement test with regard to uncertainty about previous instructions; it is aiming at a common standard that can be applied to all.

The aptitude test is an indicator of some future performance by the individual. A language aptitude test, therefore, is a prognostic measure that indicates whether a student is likely to learn a second language readily.

The diagnostic test differs from the others in that it relates to the use of information obtained and to the absence of a skill in the learner. A diagnostic test may be constructed for itself or it may be an additional use made of an achievement or proficiency test.

When to use what test is a matter of educational consideration. The user should be fully aware of his purposes before he decides on the kind of test to use.

Table 7.1. Uses of Different Types of Tests

	Achievement	Proficiency	Aptitude	Diagnostic
Feedback & Backwash	*	*		*
Selection	*	*	*	
Progress Check	*			*
Proficiency Evaluation	*			
Prediction		*	*	

2. According to the way test scores are reflected

The norm-referenced test compares a student's performance against the performance of other students. The norm-referenced test is used for the purpose of selection of candidates by setting up either percentile or standard score norms.

The criterion-referenced test indicates whether the student has met predetermined objectives or criteria. The achievement test is very often criterion-referenced because it is geared to certain known syllabuses.

3. According to the manner in which test items are graded

The objective test consists of objective items, each of which has a specific correct response. Therefore it is always scored "objectively" whether the items are scored by teachers or by machines, by one teacher or another, today or last week. Objective items include multiple-choice items as well as fill-in-the-blank and short-answer questions.

The subjective test consists of subjective items that do not have a single right answer. A subjective item may be scored in different ways by different teachers under different circumstances. Subjective items include compositions, translations, or oral interviews.

4. According to the way the skills and abilities of the students are measured

Discrete-point test measures whether or not the students have mastered specific elements of skills or abilities. In the case of a language test, items can be set to test the students' mastery of vocabulary, structures, or phonetics in separation. Most multiple-choice items are discrete-point items.

The global test, in the case of a language test, has global items which measure the students' abilities to understand and use language in context. For example, a composition can measure the students' use of vocabulary and structure as well as their fluency in written expression.

5. According to test requirements

The speed test is one in which the student works against time, to complete a given task. A typical speed test is the typing test in which the student tries to improve his or her rate of words per minutes.

The power test is one in which the student is given sufficient time to finish the test and what is being measured is the power of the student to do a task.

6. According to the dimension in which tests are used

The large-scale test is used in citywide and statewide testing programs, made practical by the advances of technology for processing the examinee responses. A large-scale test is a standardized test in the sense that it has been tried out on a proper sample of the population for whom it is intended and that on this sample it has been shown to be both reliable and valid.

The classroom test is an ad hoc test which any teacher may wish to construct and use himself in his own teaching. The most familiar role of the classroom test is to furnish an objective evaluation of each student's progress, that is, his or her attainment of course objectives and his or her performance in relation to that of the rest of the class. In general, the teacher does not have to utilize the sophisticated techniques used in large-scale testing programs.

Differentiating types of tests helps test designers visualize the character of the testing programs so that a clearly defined test blueprint can be worked out. For example, after studying the situation in China, the test constructors of the EPT decided that it should be a large-scale, norm-referenced test of English proficiency, which should consist of a large proportion of discrete-point objective items, supplemented by a subjective component. As it is a proficiency test, both power and speed should be taken into consideration. The main purpose of the test is the selection of scholars to study abroad.

Table 7.2. Appraisal of Educational Objectives of English Proficiency Test (EPT)

Bloom's Taxonomy	EPT	Percentage
Knowledge	Structure	6.3
	Vocabulary	12.5
Comprehension	Reading Comprehension	25
	Listening Comprehension	21.9
Application	Correction	6.2
Analysis	Close	12.5
Synthesis	Writing	15.6
Evaluation		

Determining the Content and Score

In this respect, Bloom's taxonomy of educational objectives (1956) may throw some light on what activities and skills to appraise in an educational test. A test does not have to cover every domain listed in the figure: however, the test constructor would be well advised to consider in every test situation the applicability of some type of classification of behavioral objectives. Even a limited application should help him or her break away from the traditional pattern of testing which is confined to knowledge of definition and facts. The taxonomy is a hierarchy of objectives representing different levels of intellectual abilities and skills. Table 7.3. shows how syntax can be tested in terms of these levels.

A rule of thumb for selecting content areas is to follow the normal distribution: having fewer items at the two ends of the taxonomy and concentrating on the middle. In Table 7.2. the EPT is cited as an example for illustration.

Selecting Item Types

Open-ended items are test questions for which students may give a variety of responses. Oral interview, compositions, content questions, and translations, for example, are all open-ended. The marker of open-ended items is often accused of being too subjective; however, through the development of an appropriate scoring procedure, it is possible to evaluate student performance with a good degree of objectivity.

Table 7.3. Testing Syntax in Terms of Bloom's Educational Objectives

Taxonomy	Item Types
Knowledge	Items to test knowledge of grammatical rules
Comprehension	Identification of grammatical errors
Application	Correction of grammatical errors
Analysis	Formation of sentences to express ideas
Synthesis	Combining sentences into a discourse
Evaluation	Judgment of style

Closed items are designed to elicit specific responses from the student. The multiple-choice item format is most often selected for standardized tests, whereas classroom teachers typically make heavy use of short-answer items, for which the teacher must prepare a scoring system in advance. When possible, the students should be told exactly how their performance is to be evaluated.

Another important feature of the test plan is the specification of the number of items to be included in the test as a whole and in various parts. Relative weights should be assigned to the various parts of the test. The purpose of the assigned weights is to assure that the various parts will contribute to the composite score in proper relation to their judged importance. Finally a two-axis chart of specification can be worked out.

Table 7.4. Distribution of Items in the English Proficiency Test (EPT) by Type and Complexity of Skill

	Type of Skill						
	Grammat. Structure	Voca-bulary	Reading Compre.	Cloze	Listening Compre.	Writing	Total
Knowledge	10	20					30
Comprehension			40		35		75
Applications	10						10
Analysis				20			20
Synthesis						25	25
Total	20	20	40	20	35	25	160
Percentage	12.5	12.5	25	12.5	21.9	15.6	100
Time Limits (minutes)	20	60		20	30	30	160

Producing a Test Syllabus

Producing a test syllabus is an essential step if the test is going to a standardized one and open to the public. It involves producing the syllabus for those taking the test and instructions for administering and scoring the test. In some cases, a specimen test is also included in the syllabus. The syllabus is something both the test-setters and the public have to follow, so that the test can be held constant and play the role of standardization. In China, very few tests have syllabuses to go with them, and the test-takers have no way of knowing in advance how thy are going to be tested. In addition, the test-users have no way of knowing how the test scores are going to be interpreted. The EPT has a syllabus and a set of specimen tests, and they are known to the public. The production of a test syllabus should be considered as a major step towards standardization.

Stage Two: Pretesting

Requirements

Pretesting is an important stage in standardized testing. A standardized test involves tens of thousands of test-takers, and it is a very often a selection decision-making process. In order to ensure the reliability and validity of the test, it is necessary to introduce the stage of pretesting, the central technique of which is item analysis.

Certain requirements for pretesting must be fulfilled before analyzing the gathered data. The sample for the pretesting must be representative, consistent, and secure.

The sample for the pretesting should be representative of the population to be tested and be selected by an efficient sampling procedure. Ideally, each student in the sample should be individually drawn from the population by simple random or stratified sampling. This question has to be considered in relation to the other requirement -- security. If we have no way of protecting the security of the test, then it is advisable to have the pretest in restricted areas. This also has disadvantages, since representation is best obtained by having a broad sampling of test-takers.

When we say the sample must be consistent, we mean that the same test format must be administered under the same circumstances. This is absolutely necessary for a new standardized test because the pretest will give us information on whether the test is adequate in terms of level of difficulty, test length, time limits, and weighting.

The pretest materials must be secure because the items will be used in the future. Actually one of the main purposes of pretesting is to get rid of weak or defective items, so that a pool of good items can be eventually built up. In a large country like China, where a nationwide test often involves millions of test-takers, security has always been a headache. It is advisable therefore to have large quantities of items pretested a few years earlier in different centers and build up an item bank. This is possible only after we have been quite sure of the format of the standardized test and we are not likely to change it shortly.

Item Analysis

Item analysis is central to pretesting. It seeks to provide data for determining, (1) the difficulty of each item; (2) the discrimination power for each item; (3) weak or defective items; (4) the correlation among the items in order to avoid overlap or bias in item selection; (5) the number of items constituting the final test and the appropriate time limits for the final test. Item analysis is an intricate process that can only be done on a computer. This can be managed on a microcomputer that is now easily accessible in China. The program for item analysis is known as GITEST II and was developed by the Guangzhou Institute of Foreign Language. Table 7.5. is a printout of an excerpt of the results of an item analysis.

Item Difficulty. The most popular measure of item analysis is p -- the proportion of correct answers on the item. So the p-value is actually a measure of how easy an item is: the larger the item proportion p, the easier the item. That is why p-value is also called facility value. The last two columns in Table 7.5., give us the p-value in terms of proportion (P+) and transformed standard score (Pd). For item 20, the key is D (marked *); 146 out of 174 students got it right, so P+ is 0.84. Pd is 9.05 obtained by the formula

$$\Delta = 13 + 4z$$

Table 7.5. Excerpt of EPT Item Analysis Printout

ITEM NO.: 20 KEY: D TEST CODE: EPTPG1 NO. OF CANDIDATES: 174

MT	MI	MA	MB	MC	MD	MO	A	B	C	D	O	Ar	Br	Cr	Dr	Or	P+	Pd
12.14	11.0	8.12	10.04	9.09	13.77	-5	5	2	21	146	0	0.09	0.02	0.3	0.46	0	0.84	9.05

ITEM NO.: 21 KEY: D TEST CODE: EPTPG1 NO. OF CANDIDATES: 174

MT	MI	MA	MB	MC	MD	MO	A	B	C	D	O	Ar	Br	Cr	Dr	Or	P+	Pd
12.14	12.34	11.19	12.64	7.83	14.48	0.42	62	18	3	91	0	0.4	0.02	0.06	0.51	0	0.52	12.77

ITEM NO.: 22 KEY: C TEST CODE: EPTPG1 NO. OF CANDIDATES: 174

MT	MI	MA	MB	MC	MD	MO	A	B	C	D	O	Ar	Br	Cr	Dr	Or	P+	Pd
12.14	12.34	8.01	10.05	13.99	9.21	0.42	12	12	139	11	0	0.23	0.13	0.51	0.16	9	0.8	9.65

z being the standard score of 0.84, which is -0.9875 here. Since Pd is far below the mean -13, it is a fairly easy item. We can also look at the figures below MT (the mean criterion score of the total) and MI (the mean criterion score of the subtest). Both 12.14 and 11 are below the mean, signifying that both the test as a whole and the subtest are somewhat easy. The ideal p-value for an item is 0.5. In a group of one hundred examinees, when all the items of the test are perfectly correlated, 0.5 means the best fifty students are distinguished from the other fifty students, and the number of discrimination between persons is 2,500 (50*50). If the P-value is 0.01, we can only distinguish one person from the other ninety-nine students, the level of discrimination being 99 (1*99). It is not that easy to obtain 0.5 for all the items of the test, so EPT set a limit). Any item from 0.3 to 0.7 is considered acceptable. Items below 0.3 and above 0.7 should be considered cautiously.

Item Distinguishing Power. A good item should be able to distinguish between students with high scores and those with low scores. In classroom testing, this is done by dividing the scripts of the students into three subgroups and then comparing the number of passes in the upper subgroup and the lower subgroup. The rationale is those from the upper subgroup should have more passes than those from the lower subgroup if the item has good powers of selection:

Table 7.6. Computing Discrimination Index for Classroom Use

N=36, U=12, L=12

Item	Upper subtotal (U)		Lower subtotal (L)		Difference (U-L)	Discrimination Index (U-L)/U
1	11	-	5	=	6	6/12 = 0.5
2	6	-	6	=	0	0/12 = 0.0
3	3	-	5	=	-2	-2/12 = -0.17

For item 1, eleven out of twelve of the upper subgroup got it right, whereas only five out of twelve of the lower subgroup passed; thus, the discrimination index is 0.5. But for item 2, six from each of the two groups got it right, so the discrimination index is 0. Item 3 is the worst item because the students from the lower subgroup did even better than those from the upper subgroup.

For large-scale testing the model is too simplistic because it does not differentiate students in the same subgroup. The use of the biserial-correlation coefficient is recommended:

$$r_{bis} = \frac{Mr - Mw}{St} \cdot \frac{p(1 - p)}{y}$$

where

Mr = mean criterion score for students choosing the right answer

Mw = mean criterion score for the other students

St = standard deviation of criterion scores for all students

p = proportion choosing the right answer

y = ordinate in the normal distribution, which divides the area under the curve in the proportions p and 1 - p

Table 7.7. Biserial-correlation Coefficients of Two Models

	Criterion (Total Scores)	Students Choosing: Wrong Answer	Right Answer		Totals
u					
p	20		2		2
p	19		2	=11	2
e	18		3		3
r	17	1	4		5
	16	1	4		5
	15	1	2		3
	14	2	2		4
l					
o	13	1	2		3
w	12	1	1	=5	2
e	11	2	1		3
r	10	3	1		4
Total		12	24		36

For classroom testing

$$D = \frac{(U - L)}{U} = 0.5$$

For large-scale testing

$Mr = 15.875$ $Mw = 12.75$ $St = 2.99$ $p = 0.67$ $y = 0.3635$

$$r_{bis} = \frac{15.875 - 12.75}{2.99} * \frac{0.67 * (1 - 0.67)}{0.3635} = 0.637$$

Table 7.7. compares the results of the two models, and we can see that the use of biserial-correlation coefficient yields a higher discrimination index. In Table 7.5. the numerals below Ar, Br, Cr, Dr, and Or denote the biserial-correlation coefficients of the students choosing the options A, B, C, and D and omitting the item: 0.09, 0.02, 0.3, 0.48, and 0. Items with biserial-correlation coefficients above 0.3 such as the key (D) and the "distractor" C are considered acceptable, whereas A and B must have "distracted" some of the better students. In general, it is sufficient to look at the biserial-correlation coefficient of the key.

Item moderation. One of the purposes of item analysis is to identify the weak or defective items. The row of numerals under A, B, C, D, and O in item 20 in Table 7.5. shows the number of students choosing the respective options, whereas the row under MA, MB, MC, MD, and MO gives the mean-criterion scores of these students. We can see then that five students with the mean-criterion score of 8.12 chose A, and two students with the mean-criterion score of 10.04 chose B. So both A and B are weak distractors and need to be modified. If we replace them with some stronger distractors, the facility value will be lower, and the item will have even stronger discrimination power.

Correlations among the items. To study the correlations, it is necessary to use a special technique known as factor analysis. We have been able to incorporate the factor analytic procedure in GITEST II. This procedure provides us first with a correlation matrix.

Table 7.8. shows the various correlations among the subtests of the EPT. The principal components of the correlation matrix were obtained and were followed by a varimax rotation to simplify their interpretation.

Table 7.8. Correlation Matrix of Different Sections Within EPT

	Grammat. Structure (I)	Vocab. & Read.Comp. (II)	Cloze (III)	Listen. Compre. (IV)	Guided Writing (V)	Total (IV)
(I)	--	0.70	0.62	0.55	0.08	0.77
(II)	0.70	--	0.66	0.67	0.25	0.92
(III)	0.62	0.66	--	0.64	0.21	0.81
(IV)	0.08	0.25	0.21	0.17	--	0.39
(V)	0.77	0.92	0.81	0.80	0.39	--

Table 7.9. shows that the components of the test can account for 63.5 percent of the variance (R^2). The first factor is comprehension (both reading and listening); the second factor is comprehension in connection with grammatical knowledge. Both reflect the receptive skills of the students. The third factor is written expression (cloze and guided writing), which is a productive skill. The factor analysis of the components conforms with the original intentions of the test-setters.

Table 7.9. Factor Loading after Varimax Rotation

	Factor I	Factor II	Factor III
Grammatical Structure	0.30	0.68	0.30
Vocab. & Read Comprehen.	0.67	0.53	0.18
Cloze	0.32	0.40	0.52
Listening Comprehension	0.78	0.20	0.26
Guided Writing	0.36	0.30	0.56
Contribution of Factor	1.39	1.03	0.77
Percentage	27.70	20.50	15.30
Total Percentage (R^2)		63.50	

Assembling the test. The procedure for assembling the test uses item-analysis data for selecting items with good discriminating power and specific difficulty levels for the final test. The EPT setters have predetermined somewhat arbitrarily the number of items for each level of difficulty.

Table 7.10. EPT Items at Each Level of Difficulty

Level	Number	Percentage	Facility
Very easy items	7	5	0.9--1.0
Easy items	20	15	0.7--0.9
Medium items	81	60	0.3--0.7
Difficult items	20	15	0.1--0.3
Very difficult items	7	5	0.0--0.3

GITEST II provides us with a summary report of all the items, so that we know what items to include in the final form.

In the last two rows of Table 7.11., we are given the number of items at each level against the intended number and also the number of items that passes the intended 0.3 level of biserial correlation. We know from the summary report that the test as a whole tends to be somewhat easy with too many very easy items. The report is followed by statistics of each subtest, so that we know what items to moderate or to replace.

Table 7.11. Tabulation of p-value and r_{bis} Coefficients of EPT Test

R / R	0-0.1	0.1-0.2	0.2-0.3	0.3-0.4	0.4-0.5	0.5-0.6	0.6-0.7	0.7-0.8	0.8-0.9	0.9-1	Total
0.1	0	2	0	0	0	0	0	0	0	0	2
0.1-0.2	0	2	1	0	0	0	0	0	0	1	4
0.2-0.3	0	1	1	3	0	1	2	4	2	0	14
0.3-0.4	0	1	3	2	2	3	3	1	10	0	25
0.4-0.5	0	0	0	4	4	4	6	2	4	0	24
0.5-0.6	0	0	0	4	3	8	4	11	2	0	32
0.6-0.7	0	0	0	4	3	3	9	4	1	0	24
0.7-0.8	0	0	0	0	3	3	1	0	0	0	7
0.8-0.9	0	0	0	0	0	2	1	0	0	0	3
0.9-1	0	0	0	0	0	0	0	0	0	0	0
Total	0	6	5	17	15	24	26	22	19	1	135

	VD 5% 7	D 15% 20	I 60% 81	E 15% 20	VE 5% 7	Total 135
Total	0	11	82	41	1	135
Pass	0	4	75	35	1	115

Table 7.12. Item Statistics of EPT's Listening Comprehension

EPTG1 Subtest Table
Lists

No. of Items: 35 (from 101 to 135)

Mean	SD	Variance	p^+	pd	R11	Rbis	Skewness	Kurtosis
20.53	6.63	43.96	0.59	12.13	0.95	0.53	-0.01	-0.77

Difficulty	Total	No. (4.3)	Items
VD	0	0	
D	1	0	
I	25	1	126
E	9	2	105.121
VE	0	0	

Here it must be stressed that items cannot be selected only on the basis of their statistical properties. It should be agreed by all concerned that the item is a good measure of the educational objectives to be tested. It is desirable perhaps to consider (1) whether the test serves the intended purpose in terms of content areas, item types, and weighting; (2) whether the test is reliable and valid in terms of facility values, discrimination power, item distribution, and scoring method; and (3) whether the test is feasible in terms of test length, time limits, and other administrative or environmental factors.

Stage Three: Administering the Test

In China, where testing and examinations have a long historical tradition, a sophisticated procedure of test administration has developed. I shall discuss here only briefly some of the relevant factors in connection with the administration of multiple-choice test forms because they are relative new in China not only to test writers but also to test administrators.

The very concept of a standardized test implies rigid control over the conditions of administration. A standardized test seeks to establish norms, but norms will be meaningful only if derived from the administration of the test under the established conditions. Standard uniform procedures are particularly important for an examination that is given nationwide. The scores of examinees will be comparable only if all supervisors and their assistants follow the same procedures and give exactly the same instructions.

Producing Test Booklets and Answer Sheets

Preparing test booklets and answer sheets for multiple-choice items involves a number of considerations:

The physical form. Apart from the technical problems of deciding page size, number of pages, use of cover page and blank pages, planning the layout of individual pages in relation to different types of items is obviously an important point for consideration. The best arrangement is that which gives the optimum combination of legibility and economy of space. In order to protect security, it is desirable to have different forms of the same test paper ready. These forms should be sorted and packed in bags in such a way that when they are used, the examinees sitting close to one another will not get the same form. This also applies to the production of answer sheets which should also have different formats. In large-scale testing the technique of machine scoring in the form of optical-scanning equipment and computers is used. It requires a carefully designed answer sheet.

Clarity of instructions. To ensure that every examinee gets the same instructions, it is advisable to put down clear and simple directions on both the test booklet and the answer sheet. It is the author's job to find those elements or characteristics of format that will be most effective in causing examiners and examinees to follow directions exactly. Use of such devices as boldface type, underlining, enlargement, contrasting colors, and encircling will be very helpful. For a multiple-choice test, special instructions should be made so that the examinees know whether they should respond to every questions, even if they do not know the correct answer. If guessing is not encouraged, then the author should also make that clear in the directions.

Arrangement of items. Items are usually arranged according to a level of difficulty, with the easiest item placed first. However, item difficulty must be approached with some caution, and the sequence of items cannot be based only on this information. Content and other editorial considerations should be taken into account. The EPT has a listening component and a writing component. The first component, though not easier, is put at the beginning the test for administrative convenience. The second component is the last subtest because it is difficult for the test-takers to control their time in a writing task.

Organizing the Testing Rooms

For multiple-choice tests, [illegible] is desirable to adopt measures to prevent prearranged cheating plans. For example, seating examinees in random order may prevent friends from communicating for any purpose during the test. Whenever possible, examinees should face in the same direction in any level seating arrangement and should be separated by a minimum of one and one-half meters. A record of the seating plan should be made in case it is needed for a future security check.

The supervisor is a very important person in the measurement process. His job is to carry out precisely the directions of the test author. The meaningfulness of the test scores will depend in no small measure on how seriously he takes his job. The training of supervisors and proctors is therefore essential in implementing the test.

Marking the Tests

If the test is entirely composed of a multiple-choice items, then the machine-scoring method can be applied. Considering the backwash effects of the multiple-choice test, essay-type items are very often included in the test. We then have to tackle the problem of marking the testing papers objectively. Basically it is a problem of achieving greater reliability. Different approaches have been suggested to improve the rating rating of essays. Two common approaches to the problem of obtaining reliability in the ratings of essays have been used. One attempts to reduce scoring variability by providing detailed guides to the rating, the other concentrates on global ratings and reduces the error by including a number of independent ratings in the total score. In the Chinese context, no matter which approach is used, it means a greater administrative load because the markers are required to assembled in different centers.

Gathering Data on Test Items

In China large-scale testing requires an enormously large number of personnel to manually gather and compute raw scores. In some cities the computer has been used to process the data. This is one area that promises to have great potential because nearly every large city has mainframes that are not fully utilized. The problem facing us is lack of a single model to process the data. It is therefore imperative to popularize modern statistical concepts, on the one hand, and adopt a few basic criteria for data processing, on the other. The criteria should take into consideration both objective and subjective tests. Although this discussion of item analysis mainly deals with objective tests, other analytical procedure should be sought to describe data on subjective test items.

Stage Four: Establishing the Norm

Evaluating the Test

Standardized tests are essentially norm-referenced tests. Establishing the norm is the final goal of standardized testing. Before we embark on the establishment of norms, we must study the final statistics once again, so that we can rest assured that the test is built on solid ground. We must find out whether the scores a normally distributed, the test as a whole is reliable and valid, and the difficulty level of the test is suitable.

Normalcy of score distribution. GITEST II provides us with a frequency distribution table of total scores.

Table 7.13. Item Statistics of EPT as a Whole
EPTPG1 Test Table

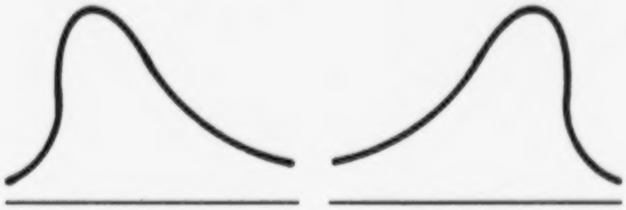
Total No. of Items: 135
Date of Test: 1983.12

Total No. of Subject: 5
Date of Analysis: 1985.6

Mean	SD	Variance	p+	pd	R11	Rbis	aVALUE	Skewness	Kurtosis
78.99	19.51	380.64	0.59	12.14	0.93	0.49	0.83	-0.13	-0.29

The skew measures the degree of departure from symmetry; the obtained value lies between -1 and +1. A positive skewness denotes that many of the items were rather difficult, so that most of the scores are gathered near the lower end. A negative skewness denotes that many of the items were relatively easy, so most of the scores are bunched at the top end.

Figure 7.2. Positive and Negative Skewness



Positive skewness

Negative skewness

Kurtosis is the sharpness of the curve of a distribution. If a distribution has a relatively high peak, then it has a positive value from 0 to 1; if a distribution has a relative flat top, then it has a negative value from 0 to -1. 0 denotes a normal distribution.

Reliability. The reliability of the test is given by Kuder-Richardson formula KR-20:

$$R = \frac{n}{n + 1} \left(1 - \frac{\sum_{i=1}^{n} p_i \, q_i}{SD} \right)$$

where n = no. of items not omitted by all candidates

SD = standard deviation of test

p_i = p-value of i item

$q_i = (1 - p_i)$

In general the reliability coefficient should be more than 0.90 if it is a large-scale test. R is also given in Table 7.12.

Validity. The validity of the test can be found by computing the correlation coefficient between the new standardized test and a well- established standardized test of similar character. When the EPT was first tried out in 1980, a group of examinees had the opportunity of taking two TOEFLs and two EPTs within a short time, so the test-setters were able to obtained a correlation matrix.

Table 7.14. Correlation Matrix Between Two EPT and Two TOEFL Tests

	EPT Sample I	EPT M1 II	TOEFL Sample III	TOEFL Dec. 1980 IV
I		0.86	0.85	0.85
II	0.86		0.86	0.86
III	0.85	0.86		0.87
IV	0.85	0.86	0.87	—

A correlation coefficient above 0.7 shows a high correlation and a marked relationship between the two elements.

Suitability of the difficulty level of the test. There are a number of reasons for deciding the difficulty level of the test. From the statistical point of view, 0.5 is considered as the ideal difficulty level.

GITEST II gives a test-evaluation report summarizing all the statistics obtained.

Table 7.15. Test Evaluation Report Produced by GITEST II

The test has a mean score of 78.99, with 59 percent of the answers correct, indicating it is a medium-level test. The standard deviation is 19.51 against the expected standard deviation of 19.92, having a range of 97, indicating the scores are well distributed. On the whole, the distribution of the scores is normal. The test is very reliable; the standard error of measurement is +-5.16.

The factor analysis of the test shows that it accounts for 64.3 percent of the knowledge and skills of the examinees. It also shows the following factors:

VOO; READ; LIST
GRAM
CLOZE

The item analysis shows 1 percent very easy items, 30 percent easy items, 60 percent items of medium level, 8 percent difficult items, and 0 percent very difficult items. On the whole, the test has good discrimination power. Twenty items do not meet the requirement.

An analysis of each component shows that:

GRAM has a mean score of 13.82, with 69 percent of the answers correct, indicating it is a medium-level component. In this component GRAM, three items do not meet the requirement.

VOO has a mean score of 11.32, with 57 percent of the answers correct, indicating it is a medium-level component. On the whole, in this component VOO, three items do not meet the requirement.

READ has a mean score of 23.18, with 58 percent of the answers correct, indicating it is a medium-level component. On the whole, in this component READ, three items do not meet the requirement.

CLOZE has a mean score of 10.14, with 51 percent of the answers correct, indicating it is a medium-level component. On the whole, in this component READ, five items do not meet the requirement.

LIST has a mean score of 20.52, with 59 percent of the answers correct, indicating it is a medium-level component. On the whole, in this component LIST, three items do not meet the requirement.

Establishing the Norm

Basically, establishing the norm means creating a frame of reference for interpreting test results. Results should be reported in units that have the following properties: (1) uniform meaning from test to test, so that a basis of comparison is provided through which we may compare different tests; (2) units of uniform size, so that a gain of ten points on one part of the scale signifies the same as a gain of ten points on any other part of the scale; (3) the characteristic measured by the test must permit the ordering of individuals along a continuum from low to high.

There are different types of norm for educational and psychological tests. Norms by age and grade, for example, describe an individual's standing in relation to other individuals who are the same age or in the same grade. The percentile rank distribution is a straightforward way of representing normative data. For example, a student with a score of 107 in the EPT falls at the 79.6th percentile, meaning that he did better than 79.6 percent of the people taking the test.

Equating Test Forms

In most standardized testing programs it is advisable to have multiple and interchangeable forms of the same test. Since two forms of a test can rarely if ever be made to be precisely equivalent in level and range of difficulty, it becomes necessary to equate the forms --to convert the system of units of one form to the system of units of the other-- so that scores derived the two forms after conversion will be directly equivalent. The last fifteen or twenty years have witnessed a vigorous development of new designs and methods for equating test scores. These new methods involve the use of the computer and a sophisticated program of controlled item pretesting and analysis.

Table 7.16. Percentile Rank Distribution Table (Scholars to Study Abroad)

Score Range	Frequency	Cumulative Frequency	Percentile Rank
155-159			
150-154			
145-149	3	3,953	100.0
140-144	6	3,950	99.9
135-139	21	3,944	99.8
130-134	58	3,923	99.2
125-129	92	3,865	97.8
120-124	123	3,773	95.4
115-119	207	3,650	92.3
110-114	295	3,443	87.1
105-109	330	3,148	79.6
100-104	385	2,181	71.3
95-99	393	2,433	61.5
90-94	354	2,000	51.6
85-89	347	1,686	42.7
80-84	301	1,339	33.9
75-79	273	1,038	26.3
70-74	198	765	19.4
65-69	160	567	14.3
60-64	125	407	10.3
55-59	74	282	7.1
50-54	68	208	5.3
45-49	57	140	3.5
40-44	83	83	2.1

References

Allen, J. P. B., and Davies, Alan, ed. Testing and Experimental Methods. Oxford University Press, 1977.

Bloom, Benjamin. Taxonomy of Educational Objectives: The Classification of Educational Goals: Handbook I. The Cognitive Domain. London: Longmans, 1956.

Thorndike, Robert. Educational Measurement. 2nd ed. Washington: American Council on Education, 1976.

Yang, Huizhong, and Gui, Shichun. "The English Proficiency Test Used in China: A Report." In New Directions in Language Testing. Edited by Lee et al. Pergamon Press, 1985.

8

ASSESSING THE QUALITY OF EDUCATION OVER TIME: THE ROLE OF THE NATIONAL ASSESSMENT OF EDUCATIONAL PROGRESS (NAEP)

Archie E. LaPointe

Testing in the United States

A Revolution in Education

The past few years have witnessed a rather significant revolution in education in the United States. A general feeling of discontent with the products of our educational system developed as policymakers at the state and federal levels addressed issues of youth employment and increasing evidence of poor performance in the basic skills by school children. Because of declining birth rates, these problems were occurring during a period of diminishing enrollment, the closing of individual school buildings, and large reductions in the number of teachers employed.

Finally, concern continued over performance on the college admission tests (the Scholastic Aptitude Test [SAT] and the American College Training Program [ACT]), which had seen significant score declines over a prolonged period.

All of these elements provided a ready context for action.

The catalyst for this revolution was Dr. Terrell H. Bell, the former Secretary of Education under President Reagan. Early in the administration of the new president, Dr. Bell called together an impressive group of lay people and scholars to consider the condition of American education. They spent a year hearing testimony and considering the elements of the problem. Their report, entitled A Nation at Risk, was to become one of a series of reports on the quality of American education that appeared within a short period of time. Six were to command considerable attention.

The theme of all these documents was that American education needed serious attention. The most significant problems were identified at the secondary-school level. American children were not being provided the quality education guaranteed them by law nor was the substance of their exposure to learning fair return for the investment of tax dollars. The theme of one report, the Paideia Proposal, was that the U.S. had met half of its legal commitments by assuring access to education to all children, and the country should now address the second issue: providing each student the same quality and the same content. Several of the criticisms had to do with the fact that performance declines were occurring during a period of time when expenditures for education were increasing at a rapid rate.

At the same time, the administration of President Reagan was attempting to move as many governmental responsibilities back to the individual states as possible. Education has traditionally been a state responsibility, and a strong tradition exists of resisting any attempt of the federal government to influence or control the curriculum or to generate nationally imposed testing programs. National, or federal, participation in education funding has always been at a very low level, usually below 10 percent. Most of the funding for education comes from local municipal governments (typically 50 percent) and the second largest portion of school funds comes from the state treasuries. So it was appropriate to focus on the states as more and more people and all sectors of society looked for reform and improvement.

Criticisms of teachers and of the education establishment was broad and general. This general disillusion with the educational establishment provided opportunities for the lay sectors of society to think about, comment upon, and, in many cases, take action on the condition of education in the nation.

This was especially true in the case of elected politicians. Many of the fifty state governors saw the concern about education as an important issue that touched the lives of large numbers of voters. State legislators as well perceived this as an important item around which they could mobilize opinion and take action. And they did.

Indeed, a year after the publication of *The Nation at Risk,* the U.S. Department of Education published a second document called *The Nation Responds.* It is a series of descriptions of what each of the fifty states has done in response to the points outlined in the initial report. This is a rather impressive registry of actions taken at several levels within states and school districts to address the concerns described in the first report. Most of the fifty states have created their own commissions of one kind or another. In some cases, single states have created several commissions to look at different aspects of the curriculum or of local institutions. Many of the states have enacted new laws requiring more stringent standards for educations and/or for secondary school graduation and several have strengthened their statewide testing or assessment procedures. Over half of the states have addressed the question of teacher qualifications and have imposed, or are attempting to impose, tests that will measure the qualifications of entering and/or practicing teachers.

Testing as a Solution

Many of these decisions involve testing in one form or another. Broad categories of questions are being asked, and often measurement is being prescribed as a response:

Q: How good or bad are local/state educational systems?
A: *Test* children and find out how well they are achieving.
Q: How effective have the reforms of the past been?
A: Compare *test* results of today to those of the past.
Q: How can we be certain students will leave school with adequate skills?
A: *Test* them at several grade levels during their education and insist on performance.
Q: How can we improve the value of a secondary certificate or diploma?
A: Award one only after students can pass a *test*.
Q: How competent are our teachers?
A: *Test* them and find out.
Q: How can we be certain that only competent people will become teachers in the system?
A: *Test* all applicants to assure that they have good skills and are competent.
Q: How will we know if the reforms being instituted now will have been effective?

A: Test now and test again in a few years to determine if they have made a difference.

Q: How does a district or state compare with others of like characteristics?

A: Use similar assessment techniques in each situation so that results will be comparable.

Many of these decisions are being turned into laws enacted by each state, and several of the states are taking the additional difficult steps of increasing taxes to support these new programs.

A set of themes underlie these events. "Competency" remains an important word both in terms of describing what children should know and be able to do as well as what teachers should bring to their profession. "Higher-order skills" is a phrase being used more and more to describe abilities to understand deeper meanings and solve problems. Scores of students on questions that measure these skills have been declining--even among the most able.

"Computer competence" and "education for a technological society" are also heard more and more as the objectives of modern education are described.

A consistent theme that runs through all of these issues is the notion of accountability: what is the nation getting for its investment in education? The states and the federal government have consistently increased the expenditures for education during the past twenty years. Local communities, states, and the federal government have invested large sums of money for general and specific programs to improve certain aspects of the system. Many of these have had to do with equity and providing basic services to specific groups of children, such as minorities, non-English-speaking students, and the handicapped. Large percentages of these budgets were devoted to improving instruction and to making materials and teachers available to improve work in the basic skills of reading, language and mathematics.

Mass-media commentators, legislators, and parents want to know what has happened as a result of these expenditures and efforts and why the current concern over the condition of education is so serious. What has the American public gotten for its money?

One immediate response to this query is to test children, teachers, and the system to determine the current status. A second answer is to continue to test over time to see if things are changing. Is the situation getting better or are declines continuing? Do trends in the data indicate that problems are growing?

Indeed, most local school districts as well as most of the fifty states have well-established and long-standing testing programs of one kind or another. Various interested parties have examined the information generated by these tests and looked for specific answers to their questions. Often they found test results lacking in specificity or they discovered that test content was inappropriate. If test results did provide information about growth or about achievement vis-à-vis specific objectives or criteria, it was often difficult to interpret those results in terms of what students should know or ought to be able to do.

Setting Educational Standards

The federal government does not have the responsibility for setting educational standards in the United States; the states are charged with that responsibility, and in most situations, they pass it along to the local communities. Local school superintendents as well as state agency officials all face the same dilemma: each community represents such a great diversity of populations, minorities, and parental interests that it is usually very difficult, if not impossible, to develop a set of standards upon which all citizens of a school district or state agree. As a result, objectives or standards are usually described in very broad, general terms.

Therefore, for example, test results that indicate that a group of children are able to achieve at certain levels with regard to certain criteria leave the reader with fundamental questions of significance: are these results acceptable? Could they be improved? Are we doing as well as other communities in the state or as well as other states in the nation? Are we as a nation doing as well as countries with whom we are competing? Inevitably it becomes important to turn these data into statements of performance that can be compared with a set of standards.

Competition is a traditional and useful technique of most Western societies. Most of the testing that has been developed in the United States relies upon comparisons with some kind of "norm." This "norm," or average, reflects what a representative sample of children or adults can do. So the notion of comparisons seems like a reasonable first step in interpreting performance statistics.

However, just as no national or standard curriculum exists in the United States, there is no national test either. No one standard is common to all states or to all school districts that administer tests. Indeed, most of these objectives or standards decisions are made at the local or state level to satisfy local sets of specifications. States and school districts want to test the content that they feel is important to their community. Therefore, they select tests from commercial test publishers that measure what they feel is important, or they create committees that develop tests that are specific to their own objectives. As a consequence, the results tend not to be comparable across school districts and across states. Indeed, many states have found it impossible for them to compare within their own boundaries the results of local school districts. American society has always prided itself on this kind of diversification. Individuality has encouraged and permitted a kind of innovation that has proven to be very useful as we have looked for new solutions to old problems.

At another level, Congress and the U.S. Department of Education, looking at the fifty states, can find no bases for comparisons. The reasons for this situation again are understandable. The characteristics of the populations within each of the fifty states, the variety of objectives of those states' educational programs, and the range in the amount of state resources devoted to providing educational services for young people are so great that comparisons would seem to make little sense. On the other hand, it can be argued that children grow up and move from state to state and the the federal government is paying some portion of the educational bill and has a responsibility to guarantee that all of its citizens are receiving a minimal educational opportunity. Added to this is an overall national concern that has to do with the quality of the human resources available to the country to achieve certain national objectives in science, health, and defense, for example. Finally, federal officials recognize that competition has traditionally been a spur to states to try harder to achieve certain objectives. Their view is that if reasonable comparisons in education can be made, they can be an effective way to encourage positive action.

Comparing Educational Achievement Among the States: "The Wall Chart"

In the absence of such available comparative data, the U.S. Department of Education created what became known as "the wall chart." This was an attempt to compare educational achievement in the fifty states using available data. These data included scores of college admissions tests, information about the amount of money spent per pupil by each of the states, and certain information about dropouts and other data that was already available. The states were ranked according to these statistics, and the results generated a great deal of debate. States found themselves on the bottom of the list for the wrong reasons. For example, scores of a state in which only 4 or 5 percent of the best of the graduating secondary school students take the SAT were compared with those of a state in which 60 or 70 percent of all students take the same test. Clearly, the data are not comparable. Similarly, the dropout rates of states with practically no minority populations were compared with those of states where significant portions of their populations do not speak English.

Reactions to the wall chart were vocal and negative. Several people suggested that some adaptation of the National Assessment of Educational Progress (NAEP) be used to permit the states to compare themselves one to another. This measurement of the "outcomes of instruction" was felt to be a more appropriate set of characteristics for comparison purposes. It has also been suggested that a set of "educational indicators" that would include achievement results and also information about the quality and experience of teaching staffs, dropouts, funding for education, and so on, would provide a broader and more accurate picture that would permit more reasonable comparisons. This debate continues and will probably not achieve resolution for some time.

In the meantime, the mosaic of testing practices that exists across the fifty states and the thousands of local districts in the country continues to be a feature of education in America. Those states that have imposed minimal-competency testing programs are discovering the strengths and weaknesses of those kinds of actions. As results are taken seriously by school boards and school administrators, teachers and administrators are preparing more carefully for the tests and, in fact, in many cases, are concentrating their instruction on those points measured by the tests. In almost all situations, however, results on these minimum competency tests do show improvement of student performance. It seems realistic to assume that the tests have had an impact on curriculum content and on teacher behavior.

Other kinds of more general tests, similar to those that have traditionally been used and are available from commercial test publishers continue to predominate. Yet a study by the Center for the Studies in Evaluation at the University of California at Los Angeles recently indicated that most test results continue to go unused by teachers and school administrators and, as a result, have little impact on school programs.

Finally, an overriding issue of concern to test publishers is the quality of their normative data. It is becoming increasingly expensive and difficult to collect sufficient information on a large enough sample of students during the test development process to generate norms that are high quality. Test publishers must ask the cooperation of school districts for which this represents a significant intrusion and a burden they are less and less willing or able to accommodate. The publishers are concerned about this problem and are working with the testing directors of the large cities and states to find solutions.

In sum, the amount of testing of American school children at the present time is greater than at any time in our history. These tests include the traditional publishers' offerings, tests developed by individual states and large school districts to reflect their own instructional objectives, and minimum competency testing programs created to monitor state programs. Participation in the national programs for college entrance tests (the SAT and the ACT) continues to grow in spite of declining total populations of young people.

Meanwhile, the lay public has become increasingly sophisticated about interpreting test results, and as a result the demand is growing for results that can be intelligently and easily interpreted and related to the questions they feel are important.

National Assessment of Educational Progress

In the United States:

- Girls continue to read better than boys.
- White boys and girls read better than black boys and girls, but the gap has been narrowing during the past ten years.
- Nine-year-old boys and girls read better today than they did ten years ago.
- Boys and girls are less able to read charts and graphs than they were ten years ago.
- Children in urban disadvantaged schools do less well in mathmematics than those attending suburban schools.
- Children whose parents graduated from secondary school read better than those whose parents did not.
- Able thirteen-year-old boys and girls do less well on higher order skills than they did five years ago.
- Boys and girls are less able to support a written proposition with logical arguments than they were five years ago.
- Children who do homework read better than those who don't.
- Fifty percent of thirteen- and seventeen-year olds no longer believe that science can help solve the world's problems of pollution, nutrition, and energy.
- Students from the southeastern part of the country have traditionally read less well then students from the other three regions. This is no longer the case.

These statements are possible because of the existence of the National Assessment of Educational Progress (NAEP).

Relationships can also be made between student achievement and teacher qualifications, school characteristics, type of instructional program, television viewing, and any other variable that may be judged significant. NAEP currently collects data on nineteen important educational issues.

NAEP was established by federal law fifteen years ago to measure what nine-, thirteen-, and seventeen-year olds and young adults know and can do in several curriculum areas (reading, writing, mathematics, science, social studies, citizenship, occupational awareness, literature, art, music, and computers). It measures performance periodically and records progress over time. NAEP uses a sophisticated sampling technique that permits it to gather robust, reliable, and valid information about the U.S.'s thirty-five million students by assessing only one hundred thousand of them using carefully prepared tests that are administered following scientifically rigorous procedures.

It measures reading every two years, mathematics and writing every four years, and other curriculum areas less frequently, depending upon national interest in a subject at any given time. It also assesses young adults between the ages of twenty-one and twenty-five to determine what percentage are literate and what percentages have achieved various levels of knowledge and skill development.

In addition to asking a sample of students to answer test questions on the subject matter, each person assessed is asked a variety of other questions:

- about themselves (sex, race, age, language)
- about their background (home, parents' education)
- about their attitudes (for example, do they like school, do they enjoy reading)
- about their ambitions (do they plan on going to college)
- how they are taught a subject (what techniques do their teachers use)

No student (in the sample assessed by NAEP) is asked to sign his or her name, so the results are anonymous. The answers to each of these questions about background, attitudes, ambitions, and so on can be related to the scores achieved in reading and mathematics, for example.

A sample of the teachers of students who are assessed is also asked a series of questions:

- about themselves
- about their training and background
- about their attitudes concerning what they teach
- about how they teach

Since teachers are linked to students through the use of anonymous code numbers, the answers to each of these questions about teaching can also be related to the achievement levels of their students.

Finally, a sample of the principals of the schools are asked questions about:

- themselves and their backgrounds
- their experiences and training
- how they spend their time
- the environment of their building

These answers are also anonymous but all can be related to the appropriate students' achievement scores.

Because of the sampling and statistical techniques used, a great deal of information can be gathered about a wide range of curriculum content and about many characteristics of students, teachers, and schools. Each student in the sample spends only one hour taking the assessments.

Each new assessment contains questions asked in previous assessments so that comparisons over time can be made. Item Response Theory (IRT) scaling is now being done where possible so that trends over time can be reported more effectively.

At first, NAEP was designed to provide only national results. It is now possible to provide results at the state level, so that states can compare themselves to the national statistics. School buildings and school districts may also sample their populations and compare their results to those of the state and of the country.

Most of the recent national reports on education, legislation, and a great deal of the media discussions about education refer to NAEP data. It has achieved a position of respect for reliability and validity. It also serves as a resource for the states and for local school districts who are encouraged to use questions from the assessments as well as adapt the methodology in preparing their own local assessments.

The Assessment Process

Differences between NAEP and Other Standardized Tests in the United States

NAEP	Tests
Measures broad knowledge and skill achievement	Curriculum specific
Questions are objective-referenced	Questions are criterion-referenced
Assesses a sample of students	Tests all students
Reports group data only	Reports individual's score
Asks easy, medium, and difficult questions	All questions of "average" difficulty
No student takes a "whole" test	Every student attempts to answer all questions
Assessment takes one hour	Testing may require three-six hours
Achievement related to background information, teacher characteristics, and school descriptors	Achievement related to national norm
Permits diagnoses of groups and types of schools	Diagnoses individual student strengths and weaknesses

While NEAP differs in some significant ways from other standardized achievement tests, many of its processes are identical to those followed for the development and use of any standardized evaluation instrument. These processes are as follows:

Setting Objectives

Since NAEP attempts to assess goals common to the diverse national population, it limits itself to identifying broad objectives rather than specifying detailed criteria to measure. For example, in arithmetic it attempts to measure children's "ability to do addition" rather than their "ability to add columns of two-digit numbers with carrying." Which objectives should be assessed? What should nine-, thirteen- and seventeen-year olds know? These decisions are made by carefully chosen committees (five to nine people) of experts, teachers and specialists. Their conclusions are then reviewed by mail by one hundred people who represent various ethnic groups and geographic regions, school administrators, teachers, parents, business people, and union representatives.

Finally consensus is reached and objectives are set for each curriculum area and age level.

Creating Test Questions

Experts then develop multiple-choice and essay (open-ended) questions designed to measure these objectives.

The attempt is made to create some questions that are easy, that is, that 70 percent or more of the students will be able to answer correctly; some that are of average difficulty; and some that are quite challenging, that is, those that fewer than 30-20 percent will be able to answer correctly.

Pretesting Questions

The only way to verify that a test question will "work," that is, that it will be understood by students, is to ask a representative sample of students (about two hundred) to answer it. This pretest sample is selected from various types of schools from across the country.

The pretests are scored and each question undergoes item analysis.

These pretest results permit test makers to select good questions for the assembly of the final assessment.

Revising and Assembling the Assessments

After the field tests identify a number of questions that seem to be effective, an assessment can be assembled that reflects the agreed-upon objectives.

Since students have a wide range of abilities, some easy, some average, and some difficult questions will be included. In addition to the newly created questions, a few questions used in previous assessments are included so that comparisons can be made over time.

By comparing the answers given by thirteen-year-olds in 1985 with those given by students of the same age in 1980, it is possible to develop trend data. Careful planning is required to anticipate the kinds of information that might be useful in the future.

Selecting a Sample of Students to Take the Assessment

Sampling procedures must be rigorously followed in order to yield results that are consistent with the data requirements. For example geographic areas may be important. Urban vs. rural data is desirable. Ethnic groups are a concern. The sample for each category of interest must be large enough and free from bias in order to yield results that are dependable and comparable.

Background and Attitude Questions

Important policy questions can always be addressed when educational personnel and students are surveyed. These might be questions about equal opportunities by minority students, quality of education, finance, use of leisure time, impact of family traditions and values, quality of teaching, the characteristics of effective schools.

If a set of these issues can be identified, questions can be asked of students, teachers, and principals that will yield data that could be useful to policymakers and to those responsible for planning and administration.

Administering the Assessment

For the final data to be useful and reliable, the administration of the assessments must be rigorously controlled. They should be given using standardized and carefully timed procedures to insure comparable data. Careful training and monitoring of personnel is essential.

Scoring

Essay and open-ended answers must be scored using reliable and uniform procedures. This process requires careful training and constant supervision of the readers and the careful monitoring of results. Multiple-choice answers can be efficiently scored using mechanical procedures and consistent verification to assure the quality of the data.

Analyzing Results

The capabilities of the modern computer and the power of the newer statistical designs generate almost limitless opportunities for analyses. In order to insure the value and efficiency of this process, the quality of the data must be assured, assumptions and procedures must be constantly checked, standard error rates must be consistently verified, and results evaluated for reasonableness as well as for meaning.

The computer and data-processing personnel required to perform these tasks must be working under the direction of very capable measurement personnel who can provide direction and evaluate the results.

Reporting Results

Such a system can only be justified if its results are used.

Results will be used if they are disseminated to the appropriate people in ways that are understandable.

Data and information from a NAEP-like system allows:

(1) government leaders to assess educational achievement at the national level and perceive trends in curriculum areas, in geographic regions, for certain subpopulations, and in specific knowledge and skill areas; policymakers at this level to measure the impact of certain programs after they have been in place for a period of time.

(2) building principals and teachers to evaluate their success compared to the results of other classrooms and/or to the school as a whole in certain curriculum subjects or skills. This would probably require administrating the assessment of 100 percent of the students, but the information may be valuable enough to justify the expense. It could also be done as an experiment in certain buildings.

(3) comparison of schools within a school district. If a large enough sample of students from each building has been given the exercises in the assessment, sufficient data may be available for this kind of analysis.

(4) comparisons of the performance and of the characteristics of different districts within a state. It would be necessary to sample broadly enough from each district.

(5) the comparisons of states by aggregating data from districts or by developing an optimum sample for the state itself.

The primary purposes of all of these comparisons are to permit local, state, and national educational leaders to:

(1) identify areas of strength that others may emulate or be challenged by; and

(2) identify problem areas or trends that require assistance or intervention.

Summary

A national assessment provides, in very cost-effective ways, reliable information about the status of the country's human resources in terms of knowledge and skill development. This data can constitute a basis for setting realistic targets for future achievement. By assessing relatively small numbers of students, teachers, and principals, enormous amounts of information can be made available to those who plan for human resources, those who set the country's educational objectives, those attempting to set curricula, those responsible for training teachers and improving the effectiveness of schools, and ultimately to parents who are concerned about their children's development.

NAEP STATE ASSESSMENT OPTIONS

- Available now for the Spring 1986 Assessment of:
  - Reading
  - Science
  - Mathematics
  - Computer Competence
- Provide achievement information representative of all students in the state
  - aged 9, 13, and 17
  - in the 3rd, 7th, and 11th grades
- Enable states to:
  - measure student knowledge
  - understand the characteristics of students, teachers, and schools
  - examine the relationship between students' achievement and their backgrounds and attitudes
  - track educational progress over time
  - make direct comparisons with the region and the nation

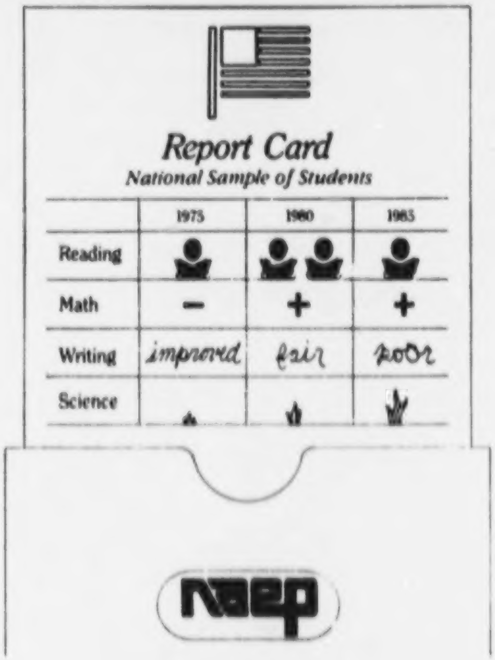
THE NATIONAL ASSESSMENT OF EDUCATIONAL PROGRESS

is a unique research project carried out cooperatively by the educational community, the states, and the federal government.

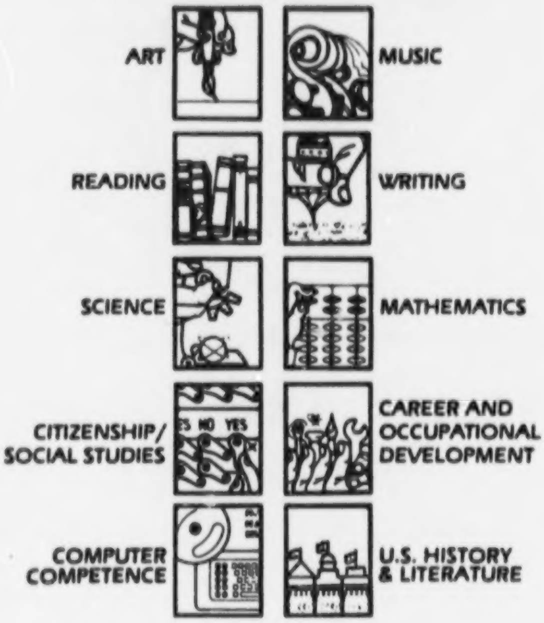
- Mandated by the Education Amendments Act of 1978
- Funded by the National Institute of Education (NIE)
- Administered by Educational Testing Service (ETS)

NAEP was developed in the mid-1960s to determine on an ongoing basis what young Americans know and can do.

THE NATIONAL ASSESSMENT OF EDUCATIONAL PROGRESS


Report Card
National Sample of Students
1975
1980
1983
Reading
Math
−
+
+
Writing
improved
fair
poor
Science
naep

LEARNING AREAS ASSESSED


ART
MUSIC
READING
WRITING
SCIENCE
MATHEMATICS
CITIZENSHIP/
SOCIAL STUDIES
CAREER AND
OCCUPATIONAL
DEVELOPMENT
COMPUTER
COMPETENCE
U.S. HISTORY
& LITERATURE

PURPOSE

National Assessment was designed to:

- MEASURE knowledge, skills, understandings and attitudes of 9-, 13- and 17-year-olds at grades 3, 7, and 11
- MONITOR performance over time
- REPORT findings to educational policy- and decision-makers
- SHARE its materials and methodology with others
- PROVIDE a national data base that can be used to address policy issues

SOME NAEP REPORTING CATEGORIES

AGE
- 9-year olds
- 13-year olds
- 17-year olds

GRADE
- 3rd
- 7th
- 11th

SEX
- Males
- Females

RACE/ETHNICITY
- White
- Black
- Hispanic
- Other

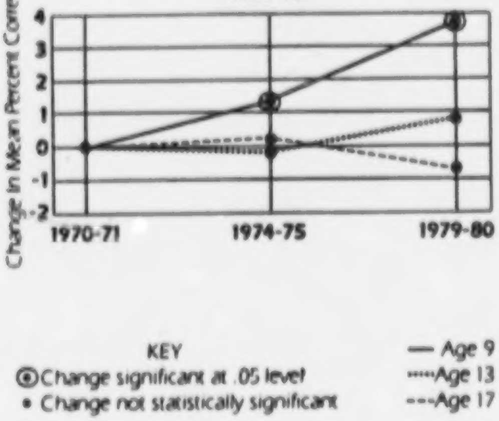
TYPES OF SCHOOLS
- Advantaged Urban
- Disadvantaged Urban

REGION OF THE COUNTRY
- Northeast
- Southeast
- Central
- West

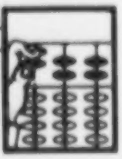
ACHIEVEMENT TRENDS

Reading

1970-80

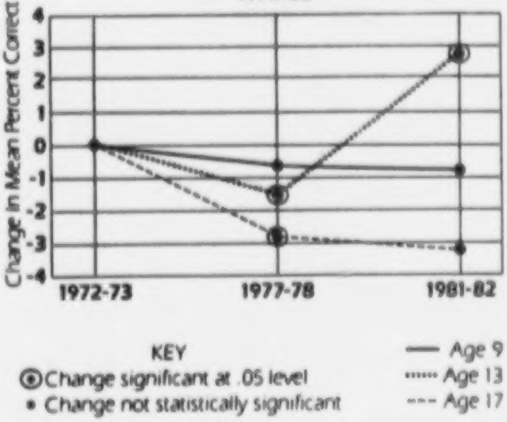


ACHIEVEMENT TRENDS

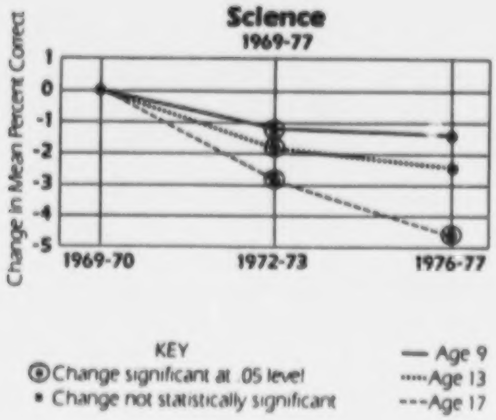


Mathematics

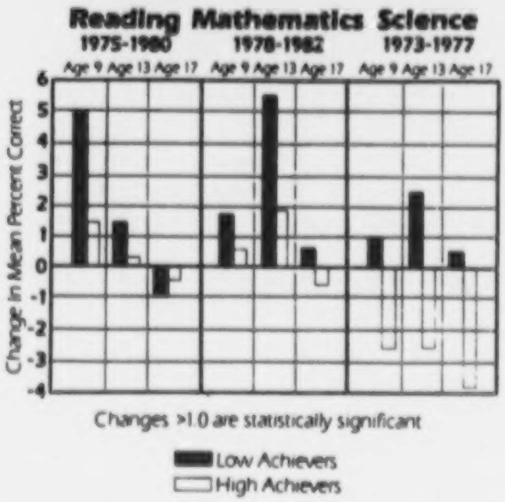
1972-82



ACHIEVEMENT TRENDS



CHANGES IN READING, MATHEMATICS & SCIENCE FOR LOW AND HIGH ACHIEVERS



Changes >1.0 are statistically significant

SOME NAEP USERS

- Local School Districts
- School Board Members
- Parents
- Teachers
- State Legislators
- State Education Officials
- Congress
- Federal Education Officials
- Education Associations
- Researchers

References

Adler, Mortimer J. The Paideia Proposal: An Educational Manifesto New York: MacMillan 1982.

Messick, Samuel; Beaton, Albert; Lord, Frederic. National Assessment of Educational Progress Reconsidered: A New Design for a New Era. Washington, D.C.: National Assessment of Educational Progress, March 1983.

A Nation at Risk: The Imperative for Educational Reform. Washington, D.C.: The National Commission on Excellence in Education, April 1983.

The Nation Responds: Recent Efforts to Improve Education. Washington, D.C.: U.S. Department of Education, May 1984.

9

CROSS-NATIONAL COMPARISONS IN EDUCATIONAL ACHIEVEMENT: THE ROLE OF THE INTERNATIONAL ASSOCIATION FOR THE EVALUATION OF EDUCATIONAL ACHIEVEMENT (IEA)

John Philip Keeves

Introduction

One of the most significant contributions of the research program of the International Association for the Evaluation of Educational Achievement (IEA) has been the change that it has wrought on the study of comparative education. IEA has achieved what it set out to do: use the similarities and differences in the provision of education in different countries to explain the forces that influence educational outcomes not only within countries but also between them. In this way IEA has made a significant contribution to the understanding of the educational processes that operate in different countries. Our purpose here is to recount how the examination of cross-national differences in educational achievement has provided firm evidence for scholars to consider in the field of comparative education (Eckstein and Noah, 1969), as well as adding important findings to the body of knowledge about education that has been assembled in the International Encyclopedia of Education (Husén and Postlethwaite, 1985).

The leaders of the IEA enterprise, which has always had a research and investigatory purpose, have repeatedly emphasized that they did not plan the studies as an Olympiad consisting of events concerned with the measurement of performance in different areas of school curriculum. Nevertheless, while it has been important for these research workers to reduce the emphasis on crude cross-national comparisons of educational achievement, these comparisons have provided firm evidence and a basis for their examination of the educational process. In order to make these comparisons they undertook studies initially in twenty-one countries, (now forty countries), which employed standardized achievement tests across key areas of the school curriculum. Furthermore, they have not limited their efforts to the measurement of cognitive outcomes but have also sought to assess the attitudes developed by students during their years of schooling. In order to provide a firm foundation for the cross-national comparisons, these research workers have endeavored to employ strictly comparable samples of students from one country to the next and to reduce the errors of sampling by the use of efficient sample designs. Moreover, they have been concerned with the accurate estimation of sampling errors, not only for mean values and univariate measures but also for bivariate and multivariate statistics. It is important to recognize that their work in the areas of test construction, attitude scale measurement, sample design, and estimation of sampling errors has been pioneering in nearly all countries engaged in the IEA enterprise.

Their program of research has been beset with frustrations and difficulties. Moreover, their work has been subjected to strong criticism, particularly from those who have rejected the use of the positivist paradigm in the field of educational research, as well as from those who have sought to defend their academic traditions from the move toward comprehensive education that has been taking place during the past thirty years (Freudenthal, 1973). Some of the key figures of IEA have recognized the shortcomings of their work with respect to its conception, administration, and methodology (Husén, 1979). However, they have in general defended the approaches taken in terms of the prevailing paradigms and the current knowledge of research methods as well as the perspectives provided by educational theories that existed at the time the studies were designed and carried out. Perhaps the greatest doubts have arisen as to whether the measuring instruments being employed could reliably and validly be used to make comparisons between national groups as well as groups of students within countries with respect to specific outcomes of education. It is commonly argued that the outcomes of education extend beyond those that can be assessed by standardized achievement tests and Likert-type attitude scales. While this contention cannot be denied, it is necessary to recognize that such achievement tests are now widely used in schools in most parts of the world and that the tests developed by IEA research workers did not limit themselves to the simple recall of knowledge but also sought to assess performance on the cognitive skills of understanding, application, and analysis.

These uncertainties with respect to test construction and attitude scale development have led some to argue, quite vehemently, that the emphasis in the reporting of IEA studies should be on individual items, the validity of which might be assessed at face value. Nevertheless, IEA has always chosen to report on total test scores that have been corrected for guessing, as has been considered appropriate across countries, as well as for performance on particular items. However, for some, test validity has been in doubt because it has been argued that no single test could be appropriate to the curricula of all countries. This question does not have a simple resolution. The appropriateness of a single test across a wide range of countries can be supported by evidence of relationships between the test and the curriculum of the schools in each country and by the estimated reliability of the test in the different countries. Nevertheless, the strength of any test must ultimately rest on its capacity, after rigorous analysis, to reveal relationships across countries that could provide an explanation of the operation of educative forces at work in the different countries. If the tests were unable to yield relationships that were meaningful in accounting for the differences between countries in their levels of performance, then it must be assumed either that the measuring instruments employed and the test lacked the necessary validity or that the hypothesized relationships did not exist. However, if the hypothesized relationships were confirmed in the cross-national analyses, then it must be accepted that, in spite of some obvious shortcomings, the tests were strong enough to serve the purposes for which they were designed.

In reviewing the IEA research studies, Inkeles (1979) has acknowledged the reasons advanced by key figures in the IEA research team for seeking to play down the Olympic Games aspect of their studies of educational achievement. Moreover, he has expressed appreciation of the sentiments which have motivated this decision. Nevertheless, he has argued that such a view was seriously mistaken since it failed to take full advantage of a unique body of data and to extract from it as much as could be meaningfully derived to provide an explanation of the cross-national differences in educational achievement and thus to obtain generalizations that might yield a greater understanding of the educational process. The current debate which has arisen in many of the more developed countries to improve the quality of education makes use of results of cross-national studies of academic achievement. Indeed, in one country at least, namely the United States, the cross-national comparisons reported by IEA have been used to show the need for a greater emphasis in schooling on educational outcomes and for a need to raise educational standards in that country (National Commission on Excellence in Education, 1983). Consequently, it is appropriate at this time to attempt to summarize what is known from the IEA studies which have been conducted and where cross-national comparisons have been carried out, with respect to providing an explanation of the differences in educational achievement between different countries.

Achievement in Reading

If consideration of reading comprehension is restricted to the levels of ten-year olds and fourteen-year olds and to performance in developed countries, then the most striking finding across the twelve developed countries involved is the small differences between them. At the high end, New Zealand had a mean score of 70 percent at the fourteen-year old level, and at the low end, Israel and Belgium (Flemish) had mean scores of 57 percent and 59 percent, respectively. However, the remaining nine countries were within two percent of the 65 percent mean score. It must be recognized that nine different languages were involved, and while the reading comprehension tests were developed in the United States, and inevitably the English language dominated the test development phase, the tests were translated effectively into a wide range of languages with little difference between countries in their mean levels of achievement. A similar pattern of results was observed at the ten-year-old level for reading comprehension tests in these developed countries. However, a similar degree of homogeniety in results was not observed for the word knowledge tests or the reading speed tests. Presumably substantial differences exist between languages in the translation of particular words across countries because uniformity in performance emerged on the word knowledge tests across the English-speaking countries involved in the study (Thorndike, 1973:23).

This study also found that little attempt was made to teach reading comprehension beyond the third grade and, as Inkeles has pointed out, these findings suggest that:

> When the student groups are basically alike in age, grade and representativeness of their respective national populations, there is very little difference between the students from one country or another within the set of MDCs (more developed countries) (Inkeles, 1979:390).

The essential skill being assessed by the reading comprehension tests was probably that of reasoning, and the inference to be drawn was that where differences were found in other subject areas between students from developed countries, such differences were likely to be associated with curricular differences in the quality of education provided.

Achievement in Science at the Lower Secondary-School Level

It would seem appropriate to use this similarity in achievement on the reading tests to examine the observed cross-national differences in science achievement. However, it is necessary to recognize that the cross-sectional nature of the studies did not readily provide a means by which the effects of the schools on educational outcomes could be completely separated from the effects of the home backgrounds of the students. Nevertheless, because testing was carried out at the two age levels for ten-year-old and fourteen-year-old students, it was still possible to examine the effects of schooling on the students in each of the developed countries taking the science tests at these two age levels.

Furthermore, because of the uniformity in reading comprehension achievement at the two age levels, as Inkeles (1979) has pointed out, it was possible to use the small differences in reading comprehension achievement to control for differences that were associated merely with increases in age, or in reasoning ability across the two age levels. These ideas were employed in a statistically simple but visually effective analysis undertaken by Coleman (1983) in a re-examination of the cross-national data available on achievement in science. The countries for which data were available fell into two groups. In one group were England, Hungary, Scotland, and Sweden; in the other group were Belgium (Flemish), Belgium (French), Finland, Italy, the Netherlands, and the United States.

The essential difference between the countries in these two groups appeared to lie in the degree of specialization introduced into the science curriculum at the lower secondary-school level or, in Coleman's words, "in the intensity with which science is taught" (Coleman, 1983: 405). Moreover, Coleman's simple graphs make it clear that the gain in science achievement relative to the gain in achievement in reading comprehension is remarkably similar within the two groups of countries and strikingly different between the two groups. Coleman also noted an increase in the spread of scores of the groups of students in those countries where specialization occurred as well as the increase in the mean level of achievement. Coleman has pointed out that major differences between school systems could have a significant effect on the achievement of students and that cross-national comparisons should be carried out in ways that would reveal these differences.

The Effects of Time Spent on Learning

In 1968 Carroll (1963) proposed a model of school learning, which included five factors that were grouped under two headings: (1) determinants of time needed for learning and (2) determinants of time spend on learning. This model lay dormant for a nearly decade. For many the importance of time was self-evident. However, the effects of time spent in learning were strongly challenged (Husén, 1974), and gradually the evidence was assembled--initially from the IEA studies of mathematics (Keeves, 1968)--to endorse the importance of the effects of time. As the data available from the IEA studies were examined it was seen that both between and within countries the different indices that were related to measures of time spent in learning were contributing to an explanation of the variability between countries, between schools, and between students in the outcomes of education. Not only was time spent in class important, but also time spent at home on study made a significant contribution. This is consistent with Coleman's view that the intensity with which a subject is taught accounted for cross-national differences in achievement.

The most convincing examination of the effects of time on learning has been presented by Carroll (1975) in his report on the study of French as a foreign language. This subject area was a particularly appropriate field for such an investigation because a large majority of the students studying French as a foreign language began learning French at school and gained little knowledge about the subject from outside the formal school setting. This is not necessarily true for other school subjects. Moreover, students would be able to report accurately when they began to learn French, information that might not be obtained without substantial error in other subject areas. In the investigation of French as a foreign language, only six countries took part at the seventeen-year-old level and eight countries at the terminal secondary-school level. To permit the examination of the effects of length of time for both population levels taken together, an achievement scale was devised that was common to both populations. Quite striking linear relationships were observed when achievement on the reading, writing, listening, and speaking tests were plotted against the number of years of French study. The countries which deviated from this general relationship were Rumania, where the students performed above expectation, and Chile, a developing country in which students performed below expectation. It was evident that the number of years of study of French was an important predictor of performance at the cross-national level. In general, analyses conducted between students in each country with respect to performance in French endorsed the view that time-related variables such as grade level and years of French study were important for achievement outcomes. However, the variable grade of beginning French gave no significant result. This would seem to contradict the assumption that an early start in studying French was an advantage.

Opportunity to Learn in Mathematics and Science

Both the mathematics and science tests were developed with great care to guarantee that they would be suitable for use across the full range of countries taking part in these two studies. Nevertheless, it was recognized that the level of performance of students in each country would depend on the nature of and the emphasis on the mathematics and science courses being taught within each country. Both the cross-national analysis by Coleman (1983) and the work of Carroll (1975) endorsed the view that the intensity of mathematics and science teaching were likely to influence achievement outcomes. In order to assess the type of mathematics and science courses being taught in each country, at about the time that testing took place, the teachers in each school were asked to consider whether the students in their school who were taking the tests had had the opportunity to learn the content of the items. By combining the responses from the teachers within a school and within a country, it was possible to assess with some accuracy the opportunity that the student within a country had to learn the content covered by the tests. Opportunity to learn was thus a measure of the operative curriculum as distinct from the prescribed curriculum laid down in a published syllabus or the achieved curriculum as assessed by student performance on a achievement test.

In the first study of achievement in mathematics (Husén, 1967), strong relationships were recorded across countries between opportunity-to-learn ratings and national mean scores on the mathematics tests. At the eigth-grade level a correlation of above 0.9 was recorded, and at the pre-university level a correlation of above 0.80 was observed. Thus it was found that students scored higher in countries where the tests were considered by their teachers to be more suited to the students' learning experiences. Strong positive correlations were recorded for a majority of countries between the total mathematics score for each school and the teachers' ratings of opportunity to learn the content tested.

No recognizable relationship existed in 1970 between the performance on the science tests of students at the ten-year-old level and the opportunity to learn the content tested as assessed by the teachers in the schools. These results suggest that at this age level the knowledge and understanding of science acquired by students in most parts of the world was gained by what was taught both in the schools and, in an informal manner, from general reading and the mass media. Consequently, the teachers' perceptions of the suitability of the test items, or perhaps the expectations of the teachers with regard to student performance, did not correspond closely with actual levels of performance.

At the fourteen-year-old level in the more developed countries, a general relationship was found to exist between level of performance on the tests and opportunity to learn the items tested. Of considerable interest was the very high level of both opportunity to learn and achievement on the science tests of Japanese students at this age level, although the level of achievement of the Japanese students was also high at the ten-year-old level. Science education in Japan differed significantly in kind from that which occurred in other parts of the world. Indeed, the evidence available showed that ten-year-old students in Japan performed at nearly the same level as fourteen-year-old students in the United States. Fourteen-year-old students in Japan exceeded, in general terms, the average level of performance of students in their final year of schooling (year 12) in the United States. The Japanese students were not tested at the year 12 level, so that no estimates could be made of their relative level of achievement at the terminal stage of schooling. However, the opportunity to learn science that had been provided by age fourteen almost inevitably guaranteed a very high level of both opportunity to learn and achievement at the terminal secondary-school level.

The high level of achievement of Japanese students, which was revealed by the IEA science and mathematics studies, has fascinated educators around the world for a decade. Shimahara (1985) has argued that the intensity of the education provided by Japanese schools was a contributing factor, although the costs in terms of tension in the lives of youth in Japan was relatively high. Both primary and secondary students in Japan attended school six days per week, with a half-day on Saturdays, for two hundred forty days a year. On average they did two hours of homework each night. The level of motivation was extremely high, and the atmosphere in the classroom promoted mastery learning of the group. Diligence, concentration, and attention to detail--characteristics leading to high achievement--were encouraged in all students, not just the most able.

At the terminal secondary-school level a surprising pattern emerged when the average level of performance was graphed against the opportunity to learn. At this level the four countries with a common tradition in science education--Australia, England, New Zealand, and Scotland--had average science scores several points above those of the other developed countries, when allowance was made for the students' assessed opportunity to learn. Next came the majority of the countries of Western Europe, with noticeably higher levels of performance in the Federal Republic of Germany and the Netherlands. These countries have similar traditions in science education as the United States.

Degree of Excellence

Coleman (1983) continued his analysis to consider the question of whether the performance of students in countries in which science was taught with a degree of intensity differed from that of students in the remaining countries, where specialization in science did not take place at the lower secondary-school level. By examining the performance in science of students in the top 1 percent of the age group, Coleman showed that the mean performance of this elite group, from which future scientists and technologists would be drawn, was higher in the countries that fostered specialization than in those which did not. Specialization in science during the secondary-school years lifted the level of achievement of the elite group of students who would continue to study science at the university level.

Comber and Keeves (1973) had undertaken a more extensive examination of the relative performance of the more able students by calculating the mean scores of the top 9, 5, and 1 percent of the age group, based on the scores of those students who remained at school to the terminal secondary-school stage. These three groups corresponded respectively to those who were likely to continue with higher education, those who were likely to take a science-type program, and those who were likely to specialize in science studies at the university. When the data collected in 1970 from the developed countries were examined in this way, the countries appeared to fall into distinct groups, within which the differences in level of performance of the top 1 percent and the top 5 percent were relatively small. These groups were the four countries with a British science education tradition, Australia, England, New Zealand, and Scotland; the Northern European countries, together with the United States of America, and the Romance-language countries. Common traditions in science education influenced the levels of performance in science of the more able students, whether or not large or small proportions of an age group remained at school at the terminal secondary-school stage.

A similar type of analysis has been carried out in the First IEA Mathematics Study reported by Husén (1967). A wide disparity in the mean level of achievement was found across the twelve countries participating in the study of the performance of students at the pre-university level who had taken courses in mathematics that would permit them to continue studying mathematics at university. Israel and England showed a high level of achievement, and Australia and the United States, a low level of achievement. However, when the performance of the top 4 percent of the age group, namely those students who were most likely to study mathematics at the university level, was considered, the wide differences across countries largely disappeared. The top 4 percent of pre-university students of mathematics in Japan and Sweden had particularly high levels of achievement. This suggests a considerable degree of specialization or intensive teaching for the most able group of students in these countries. Moreover, the evidence obtained across countries suggests that the performance of the most able did not appear to be adversely affected by increasing the proportion of the age cohort who remained at school to the terminal secondary-school stage.

Retention Rates and Mathematics Achievement

Husén (1967) reported clearly identifiable inverse relationships between the mean mathematics score for these groups of students and the percentages of the age group in the population which remained at school to the pre-university level. This held both for those who took courses in mathematics which would enable them to continue with the study of mathematics at the university level and for those who did not. However, the performance of the non-mathematics students in Japan was particularly high, and that of the non-mathematics students in Sweden particularly low. Thus, in these countries where a higher proportion of the age group have remained at school, the mean level of performance of these students is higher. This occurs irrespective of whether they have continued with the study of mathematics or not. Consequently, the apparent decline in standards over time, reported in some countries at the pre-university level, must be viewed as a consequence of the changes in retention rates over time. In searching for factors associated with higher or lower retention rates, Husén (1967) considered average income per head, industrialization, and the proportion of the group attending comprehensive schools. Only the last factor showed a strong level of association, suggesting that comprehensive schooling facilitated increased retention rates.

The relationship between retention rates and the mean level of achievement led Postlethwaite (1967) to develop a concept of "yield" as the product of retention rate and mean score for the student group. It could be considered an assessment of "how many got how far". In examining changes over time simple comparisons of mean levels of achievement are likely to be misleading because they are likely to be confounded by changes in retention rates. However, the use of the measure of yield, as undertaken by Moss (1982) in the examination of sex differences in participation and achievement in mathematics courses in Australia, would seem to provide a more valid index and more meaningful findings.

These relationships between retention rates and mean levels of achievement led Walker (Husén, 1967) to construct a model of the effects of selection on both the mean and the variance of achievement test scores of a country. The basic assumption of the model was that each country had the same distribution of mathematical ability in the complete age group and that the differences in means and variances found at the pre-university stage were a consequence of the selection procedures that operated. The model employed the truncated normal distribution and made three more specific assumptions:

(1) the distribution of scores for each country, if all in the age group had taken the tests, would be normal;

(2) these hypothetical distributions would be identical for each country;

(3) those students who remained to the pre-university stage were the best in the age group in each country.

The expected mean score and variance were calculated on the basis of these assumptions. In spite of the assumptions made, the agreement between the expected mean scores and the observed mean scores was good. However, the fit between the expected and the observed variance was less satisfactory. It would seem that a model of this kind might be useful in predicting the effects of changes in retention rates on school achievement, not only in the field of mathematics but also in other subject areas. Adams (1984) has applied this model to predict the effects of sex differences in retention rates on sex differences in performance on a scholastic aptitude test.

Retention Rates and Science Achievement

In developed countries almost the entire ten-year-old-level population was in school, but this was not so for developing countries. Similarly, at the fourteen-year-old level, most developed countries had a very high proportion of the age group in school. However, at the terminal secondary-school stage, countries had widely different holding powers. In 1970 the retention rates ranged from a high figure of 75 percent of the age group still at school in the United States to a low figure of 9 percent still at school at the pre-university level in the Federal Republic of Germany. However, the recorded estimates depended in part on how the target population was defined at the year-12 level in each country.

Comber and Keeves (1973) showed that although a general relationship existed between the mean level of achievement in science at the terminal secondary-school level and the retention rates of the school system for a country, a stronger relationship was demonstrated (-0.76) between retention rates and growth in achievement in science from the ten-year-old level to the terminal secondary-school level. Only by examining growth in achievement could some allowance be made for science learning that took place during the early years of primary schooling.

The amount of growth in science achievement between the fourteen-year-old level and the terminal secondary-school stage was also found to be related to the nature and extent of social selectivity operating. Indicators of social selectivity, based on classification of the occupation of the student's father and on the levels of education attained by the parents of the student, were used. The indices employed in the analyses were the differences between the average number of years of the father's (or mother's) education for the terminal secondary- school group and that of the father's (or mother's) education for the fourteen-year-old group. For the index of the father's occupation, the ratio of the percentage at the two levels of schooling in the professional and managerial occupational groups and the percentages in the unskilled and semi-skilled worker groups were used. Correlations between growth in achievement scores and the occupational index (0.74) and the father's education index (0.46) were recorded across the thirteen developed countries for which data were available. These data indicated that the mean gain in performance from the fourteen-year-old level to the terminal secondary-school stage increased with greater educational selectivity, and this was accompanied by greater social bias.

Achievement in Developing Countries

Four developing countries--Chile, India, Iran, and Thailand--were included in the First International Science Study conducted in 1970-71 and reported by Comber and Keeves (1973). Their lower level of economic development seemed to show up very dramatically in their performance on the science tests. The level of achievement of the students from these countries was, in general, a full student standard deviation below the performance of the students in developed countries taking part in the study.

While the relevance of these tests for students in the developing countries has been questioned, and the meaningfulness of empirical studies of educational achievement in these countries challenged, it was evident that strong support existed within many developing countries for participation in the Second International Studies of Mathematics and Science. Nevertheless, the study of reading comprehension reported by Thorndike (1973: 135) leaves little doubt that the question must be raised as to whether any more than a minimal level of literacy had been achieved in these four developing countries. This low level of reading skill clearly worked against the students' responding satisfactorily to the science achievement tests, the attitude scales, and the general questionnaires.

Inkeles (1979) has advanced four hypotheses to account for the low level of performance of students in developing countries:

(1) Lack of familiarity with the testing procedures
According to this view, the whole testing program should be dismissed as irrelevant because students in the developing countries have had little experience with multiple-choice questions and with the use of optical mark reading answer sheets. This view was supported by the evidence obtained in the study that some students were apparently responding randomly to items.

(2) Curriculum deficiency
Two additional findings support this view that students in developing countries did not have an adequate opportunity to learn the content of the science tests: ratings indicating the failure of teachers in Iran and Thailand to provide an opportunity to learn, and the relatively low values recorded at the fourteen-year-old level for Chile and India and at the terminal secondary-school level for India. However, the values of opportunity-to-learn ratings in Chile and India were not the lowest recorded, and they did not indicate that performance was expected to be at a level of one standard deviation below that of developing countries (Comber and Keeves, 1973: 159). It was evident that the teachers who provided the opportunity-to-earn ratings did not expect the tests to be as difficult for their students as they turned out to be.

(3) Poverty-caused deprivation and school resources
This view argues that the low scores of students from developing countries was a consequence of poverty within these countries and resulted in very limited resources being available in the schools. The poorer countries spent much less per pupil on education then did developed countries. The classrooms were crowded with large numbers of students and had poor facilities. In addition, the teachers in developing countries were poorly trained. However, the IEA studies were not designed to provide evidence to support or reject this view.

(4) Social deprivation
According to this view students in developing countries lacked the educational experiences provided by the home and by the community that are necessary to profit from the academically oriented education offered in the schools. The IEA studies contained little evidence of the home and community environment. Moreover the investigations were not designed as longitudinal studies that could examine the learning taking place in schools over time. Even the evidence available from the growth scores could only be used in a limited way to examine change from the ten-year-old to the fourteen-year-old levels. Consequently, little data existed to support or reject this view.

All four hypotheses were plausible. However, the First International Science Study also showed differences in achievement on the science tests of a magnitude of approximately one student standard deviation between students who were black and those who were white in the United States or between students from the southern region of Italy and those from the northern region. It seemed clear that the differences observed between the performance of students in the developing countries and those in developed countries were no different in kind from the differences in achievement between certain major sub-groups within developed countries. These differences which occurred between student groups within developed countries are commonly accounted for by poverty-based deprivation and socioeconomic disadvantage and, as a consequence, curriculum deficiencies.

One additional view that does not appear to have been adequately explored in the IEA studies was the contribution to the depression of achievement in reading and science that resulted from learning in schools in a language that differed from the mother tongue. This occurred commonly in many of the developing countries, where the national language had been decreed as the language of instruction but was not the language of communication outside the classroom. In this circumstance the students had significant handicaps to overcome before they could start to work effectively in the classroom.

Further Studies in Developing Countries

Heyneman and Loxley (1983) have brought together evidence of the effects of school quality on academic achievement in science from across twenty-nine high-income and low-income countries. Of the twenty-nine sets of data examined, eighteen were drawn from the IEA data bank associated with the First International Science Study reported by Comber and Keeves (1973). The data collected from Argentina, Bolivia, Brazil, Columbia, Mexico, Paraguay, and Peru employed test items selected from those used in the First International Science Study, and the test items used in Botswana by Leimu (1976) were adapted from IEA test battery. Only the tests used in El Salvador, Uganda, and Eqypt differed from those used in the great majority of countries. While these studies did not test identical age and grade groups in every case, as the IEA studies had done, the similarity across all samples was sufficient for meaningful comparisons. Several findings are of interest.

First, the correlation (0.55) between national per capita income and mean scores for achievement in science for twenty-five countries indicated that students in the wealthier countries performed at a higher level. Second, the correlation (0.66) between national per capita income and the proportion of variance accounted for by preschool influences (primarily home background factors) showed that the effects of the home were greatest in wealthier countries. Third, in low- and middle-income countries, the sex of the student accounted for only eighteen percent of variance. Fourth, while it was possible that in the developing countries the students from lower socioeconomic backgrounds were more rigidly selected, the evidence available from twenty-five of the twenty-nine countries appeared to indicate that this was not the explanation of the greater proportion of variance accounted for by home background variables in the wealthier countries. Finally, the correlation (-0.72) between national per capita income and the proportion of the variance in science achievement accounted for by school- and teacher-quality variables indicated that in the developing countries the impact of school- and teacher-quality on achievement in science was substantially greater than in developed countries. Furthermore, careful analyses indicated that these effects were not determined by multicollinearity in the data or other statistical artifacts. These findings were a reversal of the analyses of developed countries by Comber and Keeves (1973), which showed that home circumstances had a greater influence on achievement outcomes in science then did learning conditions in the schools.

Heyneman and his colleagues (Heyneman et al., 1981) have also reviewed studies which have examined the contribution of textbooks and reading materials in schools in developing countries with a variety of achievement and attitudinal outcomes. The key investigations used in these analyses were the IEA studies undertaken in three developing countries in 1970-71. Of the eighteen separate statistical relationships reported from ten developing countries, sixteen measured the availability of textbooks in the classroom; one, the availability of a school library; and one, the amount of time spent each day reading. Fifteen of the eighteen relationships supported the view that textbooks and reading material were necessary for effective learning in developing countries. The two negative relationships must be considered suspect because of the nature of the samples employed. The remaining relationship was not significant. Moreover, the studies from Chile, Thailand, and Malaysia indicated a greater effect of textbook possession on the achievement of students of low socioeconomic status. While not a great deal is known about the effects of textbooks on school achievement in developing countries, the evidence is consistent: they have a highly significant impact.

Further Cross-National Relationships

Passow and his colleagues systematically examined a very wide range of bivariate relationships between societal and educational indicators and the achievement outcomes measured in the IEA six subject study (Passow et al., 1976). In the main, where clear relationships were observed, they were associated with indicators that discriminated distinctly between developed countries and developing countries, for which, as has already been noted, marked differences existed in levels of achievement. These relationships generally added little to what was previously discussed involving the differences in educational provision and outcomes between developed and developing countries. Anderson (1976) has argued that the inclusion of developing with developed countries in the same study did not prove to be very informative, except perhaps to show that the examination of the educational systems of developing countries should be carried out separately. However, the work of Heyneman and Loxley (1983) has revealed quite clearly that in examining the relative contributions of home and school to the variation within countries of educational outcomes in science, it is relevant to bring both types of countries together in the same analyses. Furthermore, Anderson also concluded that:

> Among countries of the same type or "level" further search for correlates of national differences in average scores will not be a fruitful exercise, in my judgment. Indeed, we must expect that the search for national "production functions" of schools will be a long one and perhaps at the end unrewarding. It is likely that we will be able to identify a spectrum of production functions within each society, the main contrasts between societies lying in the mixture of types of functions rather than in some typical function. (Anderson, 1976: 280)

Anderson based this view on the fact that the use of the mean level of achievement by Passow and his colleagues (Passow et al., 1976) had proved largely unprofitable. It must be recognized that Passow and his colleagues were working with relatively few countries and with crude data for the societal and educational indicators that they employed, so that they could only safely use rank-order correlational procedures.

Anderson went on to suggest that the profile of achievement and the distribution of levels of achievement between schools and students within a country appeared to be more closely associated with technological levels and with other societal and educational characteristics than was the overall average achievement score for a country. Consequently, it is necessary to consider whether stronger and more meaningful relationships might be observed if more refined measures and recently developed techniques of analysis were used. Such measures and techniques should take into account Anderson's views that working only with national mean scores was likely to be unrewarding. However, even with mean scores the work of Carroll (1975) on the relationship between achievement in learning French as a foreign language and the number of years spent studying the subject has shown that important bivariate relationships do exist and can be detected when other factors do not confound such relationships.

Conclusion

The IEA research studies have been surveyed here, focusing on evidence in cross-national comparisons that might explain why achievement in one country in a particular subject field is higher or lower than that found in another country. This research has been concerned with the quality of education offered in different countries as it relates to the educational outcomes of achievement. Education has other functions to fulfill in any society in addition to the important one of fostering the learning of students in key areas of the curriculum. However, the IEA studies have resulted in greater understanding of the factors influencing differences in achievement outcomes between countries. Changes in educational practice should be made in order to improve the education provided in accordance with agreed-on goals. Within the next two to three years IEA researchers will once again have an opportunity to reexamine relationships between countries in achievement in science and thus to further an understanding of the educative process. To this end, new procedures of analysis should be employed in order to confirm or reject previous findings, to examine changes that have occurred over time, and to investigate more adequately areas of uncertainty and ambiguity in the previous studies.

References

Adams, R.J. *Sex Bias in ASAT? ACER Research Monograph No. 24*. Hawthorn, Victoria: Australian Research Council for Educational Research (ACER), 1984.

Anderson, C.A. "Interpreting National Contrasts in School Achievement." In *The IEA Six Subject Survey: An Empirical Study of Education in Twenty-One Countries*, edited by D. A. Walker, pp. 259-281. Stockholm: Almqvist and Wiksell, and New York: Halsted Press, 1976.

Carroll, J.B. "A Model of School Learning." *Teachers College Record*, 1963, 64, 723-733.

Carroll, J.B. *The Teaching of French as a Foreign Language in Eight Countries*. Stockholm: Almqvist and Wiksell, and New York: Halsted Press, 1975.

Coleman, J.S. "International Comparisons of Cognitive Achievement." *Phi Delta Kappan* 66 (6)(1983), 403-406.

Comber, L.C., and Keeves, J.P. *Science Education in Nineteen Countries*. Stockholm: Almqvist and Wiksell, and New York: Halsted Press, 1973.

Eckstein, M.A., and Noah, H.J. *Scientific Investigations in Comparative Education*. Toronto: Macmillan, 1969.

Freudenthal, H. "Pupils' Achievements Internationally Compared - the IEA." *Educational Studies in Mathematics* (6) (1973):127-86 (cf commentary by G.F. Peaker, *ibid*, (7) (1976):523-27).

Heyneman, S.P., Farrell, J.P., and Sepulveda-Stuardo, M.A. "Textbooks and Achievement in Developing Countries: What We Know." *Journal of Curriculum Studies*, 13 (3) (1981):227-246.

Heyneman, S.P., and Loxley, W.A. "The Effect of Primary-School Quality on Academic Achievement across Twenty-nine High- and Low-Income Countries." *American Journal of Sociology* 88 (6) (1983):1162-1194.

Husén, T., ed. *International Study of Achievement in Mathematics*. 2 vols. Stockholm: Almqvist and Wiksell, 1967.

Husén, T. *The Learning Society*. London: Methuen, 1974.

Husén, T. "An International Research Venture in Retrospect: The IEA Surveys." *Comparative Education Review* 23 (3) (1979):371-385.

Husén, T., and Postlethwaite, T.N., ed. *International Encyclopedia of Education*. 10 vol. Oxford: Pergamon, 1985.

Inkeles, A. "National Differences in Scholastic Performance." *Comparative Education Review* 23 (3)(1979):386-407.

Keeves, J.P. Variation in Mathematics Education in Australia. Hawthorn, Victoria: Australian Research Council for Educational Research (ACER), 1968.

Leimu, K. "A Report on the Initial Phase of the Botswana Qualitative Education Assessment." Mimeographed. Jyvaskyla: Institute for Educational Research, University of Jyvaskyla, 1976.

Moss, J.D. Towards Equality: Progress by Girls in Mathematics in Australian Secondary Schools. ACER Occasional Paper No. 16. Hawthorn, Victoria: Australian Research Council for Educational Research (ACER), 1982.

National Commission on Excellence in Education. A Nation at Risk. Washington, D.C.: U.S. Government Printing Office, 1983.

Passow, A.H.; Noah, H.J.; Eckstein, M.A.; and Mallea, J.R. An Empirical Comparative Study of Twenty-one Educational Systems. Stockholm: Almqvist and Wiksell, and New York: Halsted Press, 1976.

Postlethwaite, T.N. School Organization and Student Achievement. Stockholm Studies in Educational Psychology, 15, Stockholm: Almqvist and Wiksell, 1967.

Shimahara, N.K. "Japanese Education and Its Implications for U.S. Education." Phi Delta Kappan 66 (6) (1985):418-421.

Thorndike, R.L. Reading Comprehension in Fifteen Countries. Stockholm: Almqvist and Wiksell, and New York: Halsted Press, 1973.

III

TESTING FOR THE IMPROVEMENT OF EDUCATIONAL MANAGEMENT

10

EXAMINATIONS AS AN INSTRUMENT TO IMPROVE PEDAGOGY

Anthony Somerset

Introduction

Kenya's Certificate of Primary Education (CPE) was a major public examination in a developing country used to promote more effective teaching and learning in the classroom. The CPE terminated Kenya's basic education cycle and was the main instrument by which entrants to the secondary cycle were selected. The reform program started in 1974, but most of the main changes were introduced between 1976 and 1979. The program was directed from its inception by the Chief Examinations Officer in the Kenyan Ministry of Education, who later became Secretary to the Kenyan National Examinations Council when it was established in 1980. The CPE has now ceased to exist: the last group of candidates took the examination in 1983.[1]

Education and Examinations in Kenya

Kenya's educational system is undergoing a major transformation at the moment. Until 1984 schooling was divided into three cycles: a seven-year primary-school cycle, followed by a four-year secondary-school cycle, leading to a final two-year higher-secondary cycle. Each cycle was terminated by a national examination: the CPE after the primary cycle, the Kenya Certificate of Education (KCE) after the secondary cycle, and the Kenya Advanced Certificate of Education (KACE) after the higher secondary cycle. Starting in 1985 the three cycles were reduced to two. Also, primary education was extended from seven to eight years, to be followed by a single four-year secondary course. The CPE was replaced by an examination suited to eighth-grade rather than seventh-grade pupils, and a new examination for secondary leavers will be introduced in 1989 to replace the KCE and KACE.

---

[1] Between 1973 and 1977, while I was a research fellow at the Institute of Development Studies, University of Nairobi, I participated in the CPE project on a part-time basis. Then in 1977 I joined the Examinations Section of the Ministry of Education, where I was responsible for research into the development of examinations, including the CPE. I joined the new Kenyan National Examinations Council in 1980 and continued working for the organization until 1981, when I left Kenya. Much of this material is based on a longer report on the Kenyan examinations reform program: "Examinations Reform: The Kenya Experience" (World Bank, 1983). I should stress that the views I express here are entirely personal; they should not be taken as representing the views of the Kenyan National Examinations Council nor the Kenyan Ministry of Education.

Despite these changes, progress up Kenya's educational ladder will remain highly competitive. For many years, the number of pupils wanting to enter post-primary institutions has far exceeded the number of places available. In the late 1970s, for example, only about 13 percent of primary-school leavers could enter government-maintained secondary schools; less than 10 percent of secondary-school leavers could enter higher secondary schools, and less than 40 percent of higher secondary leavers could enter university.

The main reason why access to post-primary education is restricted is, of course, cost. Kenya spends a higher proportion of its national income on its schools than most countries. But because that income is low, funds available for education are severely restricted. At present, more than 25 percent of the annual government budget goes to education. This proportion could be increased only by diverting resources from other equally important sections--agriculture, irrigation, industrialization, and health, for example.

Because only a minority of pupils completing one educational cycle can continue into the next, the external examinations at the end of each cycle exert powerful backwash effects on the work of the schools. Competition among pupils and among schools to score the highest marks is intense. For the last two years of each cycle, the character and quality of teaching and learning is determined not so much by the official curriculum as by the questions asked in recent external examinations. Pupils spend a great deal of time answering questions from these examinations, and teachers model their own tests and examinations on the same questions. Commercial printers produce guides to help candidates prepare for these examinations. The guides contain answers to questions asked in previous years together with numerous practice papers. In rural areas especially, sales of examination guides make up a high proportion of total book sales.

The quality of these external terminating examinations is therefore crucial. The setters must aim to achieve two goals. First, they must produce an examination which will be an effective selection instrument. It must identify efficiently the candidates likely to make best use of the scarce places in the next cycle of education. Second, and no less important, they must produce an examination which will test the full spectrum of cognitive competencies which pupils should develop during the cycle which is just coming to an end. In particular, the examination should test terminal competencies, that is, the knowledge, concepts, and cognitive skills which will be of particular use to pupils who are not selected for further education, and who are therefore destined to enter the world of work. As we have seen, in every major national examination in Kenya, these candidates far outnumber those selected to continue their formal education. If the national examinations fail to test these competencies, the schools will have little incentive to teach them. The danger is that each cycle of the education system will become little more than a preparatory phase for the cycle that follows it.

It was a concern that the backwash effects of the CPE examination might be restricting the range of competencies developed in Kenya's primary schools that led to the examination reform program.

The Certificate of Primary Education Examination (CPE)

The CPE was the first major external examination in the Kenya educational system. It was a national examination. Candidates in all parts of Kenya--in the remotest desert regions in the far north, in the fertile tea- and coffee-growing areas in the center of the country, and in Nairobi, the capital city, and other major towns--all answered the same questions in the same examination papers on the same dates. It was a mass examination: in 1980, for example, there were about 328,000 candidates.

In theory, pupils took the CPE after seven years of schooling at age thirteen, but in practice, the majority of candidates were one or two years older. In rural areas especially, entry to grade 1 is often delayed beyond the official age of six years. In addition, a high proportion of pupils repeat one or more years during the primary cycle.

CPE Subjects

The examination tested competence in four papers on six school subjects. The subjects were: English, mathematics, science, geography, history, and civics. English was examined in two papers: one objective paper consisting of fifty multiple-choice items, which tested mainly comprehension, verbal reasoning, and knowledge of grammar and syntax; and a composition paper which tested pupils' ability to communicate effectively using written prose. Mathematics was examined in a single fifty-item multiple-choice paper. The remaining four subjects were grouped together in a single paper known as the General Paper, which was also multiple-choice in format. Forty of the items were devoted to science, twenty-five to geography, twenty to history, and five to civics. Four alternative answers (options) were provided for all multiple-choice items. Candidates recorded their answers on a specially-prepared answer sheet, which was then sensed optically by a document reader and processed by computer. The compositions were marked manually. More than one thousand examiners, mainly primary-school English teachers, were employed at four marking centers. They worked in teams of two. Each examiner in the team read the script separately. They then compared their assessments and agreed on a final mark. If their assessments differed widely, a moderator (who worked with about ten teams) read the script for a third time. No examiner was responsible for marking scripts from his home province.

All CPE subjects were examined in English, which is the medium of instruction in Kenyan schools from grade 4 onwards. The use of English gave a considerable advantage to pupils from more privileged socioeconomic groups, who frequently use English for everyday communication. However, the new grade 8 leaving examination, which will replace the CPE, will include two papers in Kiswahili, the national language.

The Use of Multiple-choice Questions

The use of multiple-choice format for all papers except English composition was dictated by necessity. Only four weeks were available for processing between the time when candidates finished taking the examination and the time when results were needed for secondary-school selection. The task would have been quite impossible with more than one open-response paper. Even so, deadlines were tight. During the initial data-capture period, when pupils' answers were being recorded on computer tape, the document reader was run continuously, twenty-four hours per day. The examinations staff worked eighty hours per week or longer.

If more time had been available, a more balanced mixture of essay-type, structured short-answer, and multiple-choice questions would have been used for the examination. Multiple-choice examinations bring substantial administrative benefits: they are quick and easy to process, and costs are reduced. They also have some professional advantages: marking is entirely objective, and a well-set multiple-choice paper can provide more thorough coverage of the subject syllabus than a paper consisting entirely of essay-type questions.

Set against these benefits are substantial professional costs. In the first place, multiple-choice examinations cannot provide adequate measures of several crucial cognitive skills, including the ability to communicate effectively through the medium of the written word and the ability to produce imaginative and original ideas. If these skills are not tested by the final terminating examination, the schools will have no incentive to develop them. In 1966 open-response questions were eliminated from the Kenya CPE, and the examination became entirely multiple-choice in format. It soon became apparent, however, that the backwash effects of this change were quite unacceptable. In most schools, the teaching of the skills of continuous prose writing fell rapidly into decay because these skills were no longer needed to score high marks. In consequence, a composition paper was restored to the examination in 1973.

Another problem is that multiple-choice questions which test the higher-level cognitive skills --"thinking" rather than "remembering" skills-- are difficult and time-consuming to construct. A good question which tests problem-solving skills or the ability to apply knowledge or concepts to new situations usually requires teamwork. The first draft may be written by a single person working alone, but to produce an acceptable final version he will need comments and criticisms from professional colleagues. A multiple-choice paper which has been written too quickly, or written by item-writers who lack the necessary experience, will nearly always contain too many questions testing straightforward recall of memorized facts.

A third limitation is more subtle. In answering a multiple-choice question involving thinking skills, the pupil can follow two strategies. He can either work from the given information to a possible answer, and then check this solution against the range of solutions offered. Alternatively, he can work backward from the given solutions, eliminating those that seem unlikely, and then perhaps guess among those remaining. A high proportion of students employ this latter strategy in answering multiple-choice questions. In the real world, this strategy is rarely of much use. Away from the classroom, cognitive skills are used mainly in "open-ended" situations. A person with a decision to make or a problem to solve has no one to specify for him in advance four or five possible solutions, one of which must be correct. Instead, he must devise a solution for himself. Thus using thinking skills in answering multiple-choice questions is essentially different from the way they are used in real-life situations.

Standardization of Scores

Before the CPE results were issued, the raw marks gained by candidates were transformed into standard scores. For each of the three main subject areas, (mathematics, English, and general subjects) the marks were converted into a standard distribution with an arbitrary mean of fifty points and an arbitrary standard deviation of fifteen points. This meant that the range of possible standard scores in each subject area was from approximately five to ninety-five (that is, three standard deviations to either side of the mean score), although occasional individuals scored outside these limits. Any standard scores of one hundred or over were brought down to ninety-nine, and any scores of zero or below brought up to one. The three scores were then summed for each candidate: the total standard score was the main criterion on which selection for secondary school was based.

It should be stressed that the values of fifty and fifteen chosen, respectively, for the mean and standard deviation of the CPE standard score distribution are essentially arbitrary. Other standard score systems have other values: the T score distribution, for example, has the same mean (fifty), but a standard deviation of only ten.

Standard scores have several advantages over raw marks. A standard score conveys definite information as to the relative performance of a candidate or group of candidates, irrespective of how easy or how difficult the questions in the examination were. If, for example, we know that a school or a district averaged sixty-five standard-score points in the CPE mathematics paper, we would know that that school or district had performed one standard deviation better than the national average in mathematics. Similarly, standard scores can be compared from subject to subject, or from year to year. A school that averages sixty standard-score points in mathematics in one year and sixty-five points in the following year has improved its mathematics performance by one-third of a standard deviation, relative to all schools in the country.

Standardization is especially desirable if marks from several papers are to be combined to give a total mark. The contribution of each mark to a total mark is determined by its scatter, not, as is sometimes thought, by the mean mark or the total mark possible. Because standardization equalizes the standard deviations of the different papers, it also equalizes their contributions to the total score.

But standard scores are measures of relative performance only; they tell us nothing about the absolute levels of achievement of the pupils sitting the examination. As we have seen, a low mean raw mark is brought up to the standard mean score, and a high mean raw mark is brought down to the same standard mean. Standard scores can thus be highly misleading if they are interpreted without reference to the raw marks from which they were derived. If, for example, most candidates gain low raw marks in an examination, standard scores will give a falsely optimistic impression of their level of performance and of the quality of the schools they have been attending.

For this reason, both standard scores and raw marks are needed in examination analysis. Here we shall use standard scores to compare the performance of groups of pupils from year to year, and raw marks to discuss the levels of mastery achieved by pupils in particular examination questions.

Instruments and Goals for Examination Reform

Educational and Allocational Goals

The reform program for the CPE examination started in 1974 and gathered momentum between 1975 and 1977. The program was directed towards five main goals, three of which can be classified as educational goals and two as allocational goals. The educational goals concern the backwash effects of the examination on teaching and learning in the primary schools; the allocation goals pertain to the functions of the examination as an allocator of secondary-school opportunities.

(1) Educational goals
  (a) Relevance
  (b) Quality
  (c) Distribution of quality

(2) Allocation Goals
  (a) Equity
  (b) Reliability

Towards these five goals, two major instruments of reform were employed: first, changes in the content of the examination papers; and second, the introduction of an information-feedback system. It was hoped that changes in the questions set would make the CPE more relevant as a leaving examination, more equitable to pupils in less-privileged socioeconomic groups, and more reliable as a selection instrument. The introduction of an information-feedback system would, it was hoped, do something to improve the overall quality of the primary-school system and to reduce quality differences between the stronger and the weaker schools.

Changes in the Content of CPE Examination Papers

Until the early 1970s most CPE questions, except in mathematics, tested little more than the candidates' ability to recall factual material. Moreover, the facts tested were mainly isolated pieces of knowledge: in history and geography, names, dates and places; in science, definitions of technical terms; and in English, grammatical rules, spelling, and use of idioms. A high proportion of the questions were similar to, or even identical with, questions which had been asked in previous CPE examinations. This approach gave a big advantage to the candidate who was prepared to spend long hours memorizing facts from his CPE guidebook and from old examination papers.

Moreover, the content of the questions was influenced much more by the function of the CPE as a secondary-school selection instrument than by its function as a primary-school leaving examination. Much of the knowledge tested was more appropriate to the junior-secondary school than to the upper-primary school; very little of it was of any practical usefulness to the large majority of pupils who left school at the end of the primary cycle.

After the reforms the examination tested a much broader spectrum of cognitive skills. The aim was to encourage teachers to develop transferable skills, that is, skills which could be applied in a wide range of contexts, both in and out of school. Many questions, for example, tested decision-making and problem-solving: pupils were expected to understand and interpret new information which they had never seen before and then to draw valid conclusions from it. Other questions indirectly tested observational and experimental skills: they gave a big advantage to pupils who had carried out practical investigations, using the resources of the local environment. The composition paper aimed to test "divergent" as well as "convergent" skills: pupils were expected to write prose which was not only grammatically accurate but also fluent and imaginative.

Remembered knowledge was still needed to answer many questions, of course, but three main changes were introduced in the testing of knowledge. First, many of the questions focused on cognitive structures rather than on single, isolated facts. They sought understanding of the relationships among facts, from which they derive their meaning, and insight into causes, reasons, and consequences. They thus tended to ask "why" and "how" more often than "what", "who", "when" and "where". The second change, closely related to the first, was that many questions required candidates to apply knowledge or cognitive structures in new, unfamiliar situations. And finally, a high proportion of the questions tested "survival" knowledge; that is, knowledge likely to be especially useful to the many pupils who would not win a secondary school place, and who would therefore be unlikely to find a formal job. This change was especially evident in science and in mathematics: much stress was given to topics such as nutrition, child care, disease prevention, first aid, energy sources, improved seeds and fertilizers, soil conservation, weighing and measuring, and the use of money.

These changes can be illustrated by some examples from science--the subject in which innovation was perhaps most radical. The syllabus of 1967 set out the main objectives for teaching this subject at the primary level:

> Lessons should be based on children's observations. One of the teacher's tasks is to put the children into situations where they can observe. Another is to try to demonstrate the principles underlying what the children see. A third is to help the children record what they see, and begin to draw conclusions from it.
>
> (Kenya Primary School Syllabus, 1967, p.111)

It will be seen that the syllabus stresses the development both of concepts and process skills. Through their experience of the world around them, pupils are to be helped to acquire an understanding of scientific principles and to develop the skills of observation, data recording and interpretation, and scientific reasoning.

The competencies tested by the sets of eight consecutive questions from the 1972 and 1973 CPE science papers are clearly quite different from those highlighted by the syllabus.

Among the sixteen questions quoted, as many as fourteen are pure knowledge questions, involving the straightforward recall of remembered facts. Only two questions, nos. 49 and 51, involve cognitive skills--in both cases, numerical calculation. Moreover, the questions focus narrowly on specific factual material, and especially on terminology. The concern is with the vocabulary of science, rather than with its ideas, its explanations, and its applications to everyday life. Moreover, many of the terms are technical (sedimentary, saline, capillarity), and some refer to pieces of equipment which are never seen in a primary school, except perhaps in some of the most privileged urban schools (barometer, telescope, hygrometer).

The 1972 and 1973 science questions reflect a preoccupation with the secondary-school selection functions of the CPE, virtually to the exclusion of its functions as a primary-school leaving examination. But even for secondary-school entrants the relevance of much of the knowledge tested must be questioned. Secondary-school pupils need, of course, to know many specific facts, including definitions of technical terms, but unless they also develop an understanding of scientific concepts, which link facts together in patterns of cause-and-effect, their store of information will be of limited usefulness. Nor is it necessary that they start learning secondary-level scientific vocabulary before they arrive at secondary school.

A selection of knowledge-based questions from the CPE science papers from 1975 to 1981 reveals several differences between these questions and those asked in 1972 and 1973. In the first place, the use of terminology was sharply reduced. The only technical terms employed were those referring to common diseases--kwashiorkor, cholera, diarrhea, bilharzia--which are the cause of much ill-health and many deaths in Kenya. Simpler words to refer to these diseases are not available in the English language. A second change is more important:

180

CPE Science Questions, 1972-1973

1972

49 The distance between the 0°C point and the 100°C point in a (mercury) Centigrade thermometer is 15 cm. If the mercury reaches a point 4·5 cm. up this scale the temperature is

A. 4·5°C.
B. $33\frac{1}{3}$°C.
C. 30°C.
D. 3°C.

50 A common cause of soil loss is

A. exhaustion.
B. erosion.
C. porosity.
D. capillarity.

51 When a substance such as magnesium is burned in a fixed volume of air, the active part of air is used up. If the percentage (by volume) of the air which is active is 20%, the volume of air left after magnesium is burned in 200 c.c. of air is

A. 140 c.c.
B. 180 c.c.
C. 160 c.c.
D. 40 c.c.

52 Rocks formed from the wearing away of existing rocks are called

A. igneous.
B. metamorphic.
C. sedimentary.
D. volcanic.

53 A telescope is used for

A. taking pictures of objects.
B. observing bacteria.
C. projecting pictures on a screen.
D. observing distant objects.

54 Soils which are found near lakes such as Rudolf and Magadi are usually

A. calcareous.
B. volcanic.
C. alluvial.
D. saline.

55 Most substances expand on heating. One substance which does not expand on heating is

A. copper.
B. glass.
C. ice.
D. iron.

56 A natural rock containing one or more metals is called

A. a mineral.
B. an ore.
C. gravel.
D. sand.

1973

41 To measure the atmospheric pressure we would use

A. a thermometer.
B. a barometer.
C. a hygrometer.
D. an anemometer.

42 Impurities suspended in water can be removed by

A. evaporation.
B. distillation.
C. condensation.
D. filtration.

43 In a flowering plant, fertilisation takes place in the

A. stigma.
B. style.
C. anther.
D. ovary.

44 Bees are referred to as social insects because

A. the queen lays many eggs at a time.
B. they make honey.
C. they share their work.
D. they live in hives.

45 Substances that allow electricity to flow through them are called

A. insulators.
B. conductors.
C. non-metals.
D. batteries.

46 A boy is able to see his image in a mirror because the light is

A. magnified.
B. reflected.
C. refracted.
D. absorbed.

47 A man who studies the stars uses a

A. telescope.
B. microscope.
C. pin-hole camera.
D. periscope.

48 Light enters our eyes through the

A. iris.
B. retina.
C. pupil.
D. lens.

Knowledge-based CPE Science Questions, 1975-1981

1975 Good farmers dip or spray their cattle to kill cattle ticks. The main reason they do this is because

A. Ticks make holes in the skin of cattle and this reduces the value of the leather.
B. Ticks feed on the blood of cattle, which makes them weak.
C. Cattle ticks carry diseases which are dangerous to human beings.
D. Ticks carry diseases which can kill cattle.

(No. 49)

1977 There is a swamp at the far end of Muturi school compound. Kithui's home is on the other side of the swamp. If he walks through this swamp on his way to school each day, which one of the following diseases is he most likely to get?

A. Bilharzia
B. Kwashiorkor
C. Rickets
D. Smallpox

(No. 48)

1977 Kushe's mother gave four reasons why she wanted to replace the old grass roof on her house with a new mabati (corrugated iron) roof. Three of her reasons are correct. Which one is WRONG?

A. A mabati roof keeps the house cooler.
B. Snakes and rats cannot hide in mabati roofs easily.
C. Clean water can be collected from the mabati roof.
D. Mabati roofs last longer than grass roofs.

(No. 53)

1977 Less than 3% of Kenya's land area is covered with forests. Four possible reasons why parts of Kenya should be kept for forests are given below. Only three of the reasons are correct. Pick out the one which is NOT correct.

A. Forests help prevent soil erosion.
B. Forests help the soil to store water.
C. Forests provide raw materials for paper manufacture.
D. Forests help prevent the spread of sleeping sickness.

(No. 54)

1977 Hadija put clay around the sides of her jiko, leaving the holes open, and let it dry. This made the jiko work much better because it

A. made the jiko heavier.
B. allowed wood to be burned in the jiko.
C. reduced the loss of heat from the sides.
D. increased the flow of the air to the jiko.

(No. 59)

1977 Wanjiku works in a health clinic. A mother brings a small child suffering from kwashiorkor. The mother says she cannot afford to buy meat, eggs, or milk for her child. What should Wanjiku advise her to do?

A. Give the child bananas or oranges from her shamba to eat each day.
B. Bring the child to the clinic each week for injections.
C. Feed the child with plenty of posho or ugali.
C. Use more beans or groundnuts in preparing the child's food.

(No. 68)

1978 There was an outbreak of foot-and-mouth disease on Katuko's farm. She was told not to move her cattle from her farm. The main reason for this was that

A. It would be easier to sell cattle on the farm.
B. Sick cattle would die if moved.
C. The cattle would become healthy if they remained on the farm.
D. Moving the cattle off the farm would spread the disease.

(No. 80)

1979 Cut dry grass is often placed between rows of coffee trees. Three of the following are reasons for doing this. Which one is NOT a reason?

A. The grass reduces the speed with which the rain strikes the soil.
B. Water evaporates from the soil more slowly.
C. The grass protects the coffee beans from pests.
D. The grass helps to prevent weeds from growing around the coffee trees.

(No. 56)

1979 Diarrhoea is a disease that kills many babies in Kenya. When babies have diarrhoea they lose a lot of water and foods. One correct way to treat babies with diarrhoea is to

A. keep them wrapped up and warm so that they sweat out the sickness.
B. give them drinks of boiled cold water containing some sugar and a little salt.
C. give them solid foods containing plenty of carbohydrates.
D. give them very little food or water until the diarrhoea stops. (No. 78)

1979 Wangari lives in a village with many people. From which one of the following sources can Wangari collect the best drinking water?

A. A big river nearby.
B. The rain from her mabati roof.
C. The dam near the village.
D. A swamp on her farm. (No. 89)

1980 Three of the following help to prevent the spread of cholera. Which one does NOT?

A. Using pit latrines or toilets.
B. Putting oil on stagnant water.
C. Washing hands before eating.
D. Keeping foods covered. (No. 53)

Which one of the following should be done immediately if a child burns her hand on a jiko?

A. Put the child's hand in cold water.
B. Burst the blister to let the liquid out.
C. Wrap the hand tightly in a bandage.
D. Cover the burn with cooking fat. (No. 89)

When Njeri was cooking mandazi on a jiko, the hot oil in the sufuria caught fire. Which one of the following should Njeri do to put out the fire?

A. Take the sufuria off the fire.
B. Put a big cover over the sufuria.
C. Pour water on the sufuria.
D. Blow very hard on the flames. (No. 90)

1981 A Standard VII class suggested the following reasons why mothers should feed their babies with breast milk rather than with bottled milk:
(i) Breast milk is easier for young babies to digest.
(ii) Babies who drink bottled milk are more likely to suffer from diarrhoea.
(iii) Bottled milk does not contain enough water for young babies.
(iv) It is cheaper to feed a baby with breast milk.
Which of these are CORRECT?

A. (i) and (ii) only.
B. (iii) and (iv) only.
C. (i), (ii), and (iv) only.
D. (i), (ii), and (iii) only. (No. 77)

Nkangi wants to plant maize and beans, but has not enough money to buy fertiliser. Which one of the following will be LEAST useful in making his soil more fertile?

A. Animal manure
B. Compost
C. Ash from fires
D. Charcoal (No. 79)

a high proportion of the knowledge-based questions tested information which is of direct application away from school in improving the quality of life. Particular attention was paid to testing knowledge which might help low-income families, in both rural and urban areas, to use their limited resources more effectively. Examples include no. 59 from the 1977 paper (cooking stoves are more efficient and burn less fuel if the metal sides are insulated with clay); no. 68 from the same paper (low-cost plant foods such as beans and groundnuts can be used to treat kwashiorkor if animal products are too expensive); and no. 77 from 1981 (it is cheaper as well as better to breast-feed babies than to bottle-feed them). A third major change is that many of the knowledge-based questions tested understanding of simple cause-and-effect concepts rather than specific facts. Pupils were asked to explain why cattle should be sprayed against ticks; why babies should be breast-fed rather than bottle-fed; why forests need to be preserved. Schoolteachers who understand the reasons for various child-care, public health, and conservation measures are likely to be more willing to carry them out.

Starting in 1974 a number of skills-based questions were introduced into the CPE science paper to supplement the knowledge-based questions. Skills-based questions involve cognitive operations. The candidates must work through a sequence of mental steps to arrive at the correct answer. The information he works with may either come from memory or be given as a part of the question. By 1978 skills-based questions predominated over knowledge-based questions in science and geography.

Following Bloom's widely used taxonomy, skills questions can be categorized into three main types: comprehension, application, and analysis.

Comprehension questions involve the understanding and interpretation of new information. The mental operations involved are essentially routine: usually, they are familiar to the pupil from previous practice. Questions no. 28 and 81 would be classified as comprehension items: no. 28 requires pupils to calculate a distance on the map from the scale; no. 81, the temperature of water from the graph.

Application and analysis questions are more complex. The sequence of mental steps to be carried out is longer and is unlikely to be familiar to the pupil from previous practice. The pupil must first decide for himself what steps are needed, and in what order, to reach a valid conclusion. Thus, questions of these types involve decision-making and problem-solving skills. Essentially, they test the pupil's ability to think effectively.

The distinction between application and analysis questions is not always easy to draw. Application questions involve the transfer of existing knowledge or concepts to new situations to solve a problem or reach a valid conclusion. Question no. 63 would be classified as an application item. A candidate is unlikely to be able to answer it correctly if he is not familiar with the concept of food chains. But he must apply this knowledge to a specific food chain--one that he has almost certainly never encountered before.

Question no. 61 would be classified as an analysis item. Like no. 63, it involves reasoning to solve a problem. But the question assumes that candidates have little or no knowledge about acids and bases: all the information needed to answer correctly is provided in the stem (the introductory part of the question). The candidate must analyze this information, search for patterns in it, and identify a valid conclusion which can be drawn from these patterns.

The distinction between application and analysis questions is not especially important. Both are essentially "thinking" questions. Their common characteristic is that the pupil must decide for himself on an appropriate sequence of mental steps to solve a problem and then carry them out accurately.

193

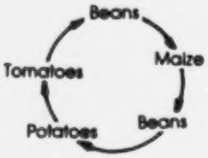
Skill-based CPE Questions, 1978-1981

60. Njeri set off to the market just after sunrise. Her shadow was strectched out to her left. Towards which direction was Njeri walking?

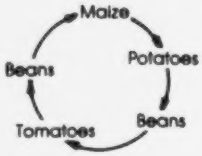
A. South.
B. North.
C. East.
D. West.

74. Beans add nitrogen to the soil which is needed by tomatoes, maize and potatoes. Tomatoes and potatoes are attacked by the same diseases which can live in the soil from one year to the next. Kamau planned to rotate the crops growing in his shamba. Which one of the following would be the best plan to use?

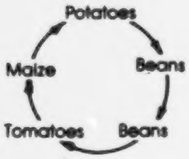
A


Beans
Maize
Beans
Potatoes
Tomatoes

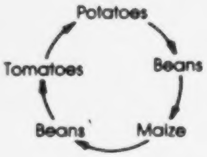
B


Maize
Potatoes
Beans
Tomatoes
Beans

C

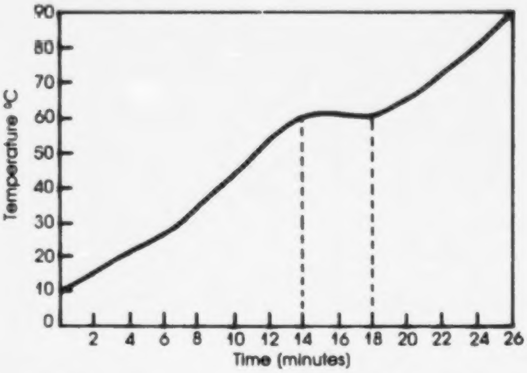

Potatoes
Beans
Beans
Tomatoes
Maize

D


Potatoes
Beans
Maize
Beans
Tomatoes

Use the following information to answer Questions 81 and 82.

Standard 7 pupils heated water in a container and recorded the temperature after every two minutes. From their readings they drew the following graph:



81. What was the temperature of the water when the heating started?

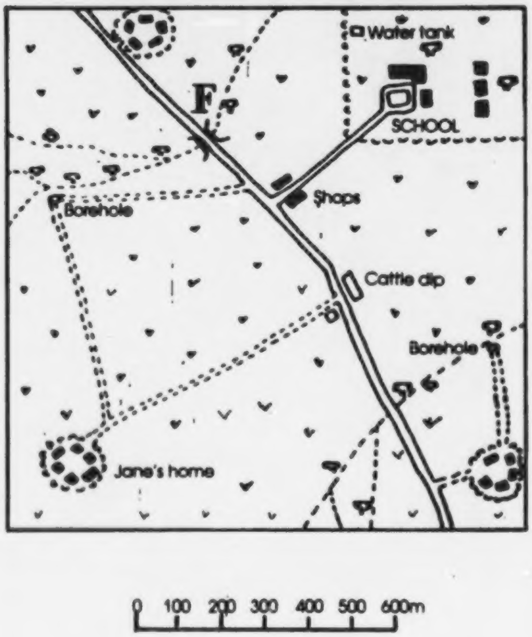
A. 0°C.
B. 10°C.
C. 26°C.
D. 90°C.

82. Which one of the following is a possible explanation for what happened between the fourteenth and the eighteenth minutes?

A. The water was boiling.
B. Heating stopped.
C. The water expanded.
D. Some water was removed.

Use the sketch map of Jane's home area drawn below to answer Questions 28 to 31.

JANE'S SCHOOL AND HOME AREA



KEY:

▬	Building	)(	Bridge
═	Road	▽	Thorn tree
⋯	Footpath	▾	Grass
⌣⌣⌣	Thorn fence	- - -	Seasonal rivers

28. Use the scale to find the shortest distance between the entrance to Jane's home and the cattle dip. The distance is about
    - A. 70 metres
    - B. 500 metres
    - C. 700 metres
    - D. 1200 metres

29. Jane goes for a walk one day. She walks north-east along the path from her home and turns right at the main road. Then she takes a path to the left. Jane will arrive at
    - A. the school
    - B. the borehole to the north of her home
    - C. the home in the south-east of the map
    - D. the bridge marked F.

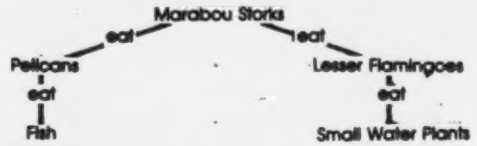
61. Substances which turn the juice from red hibiscus flowers green are called bases. Substances which turn the hibiscus juice pink are called acids. Ali tested different substances with the hibiscus juice and recorded his results in the following chart:

	lemon juice	solution of ashes	vinegar	milk	solution of soap
turns pink	yes	no	yes	no	no
turns green	no	yes	no	no	yes

Using the chart, which one of the following conclusions is correct?

- A. Milk and soap solution are both acids.
- B. Soap solution is a base and lemon juice is an acid.
- C. Ashes solution and vinegar are both bases.
- D. Vinegar is an acid and lemon juice is a base.

63. At Lake Nakuru, the following food chain can be observed.



If the lesser flamingoes are all killed, which one of the following is most likely to happen?

- A. Marabou storks will eat the small water plants.
- B. The small water plants will die.
- C. There will be fewer fish.
- D. Marabou storks will eat more pelicans.

78. Scientists at Muguga wanted to find out if they could increase the yield of sorghum, sweet potatoes and cassava by using farmyard manure. The results of their work are given below.

	Weight of the crop per hectare		
	Sorghum	Sweet potatoes	Cassava
With farmyard manure	1700 kg.	5600 kg.	9700 kg.
Without farmyard manure	1100 kg.	2600 kg.	6400 kg.

Three of the following statements about their results are true. Which one is FALSE?

- A. The cassava crop was 3300 kg per hectare heavier when manure was used.
- B. When no manure was used, sweet potatoes yielded 3000 kg per hectare less than they did with manure.
- C. The yield of all three crops increased by at least 500 kg per hectare with manure.
- D. With manure, the sweet potato crop was heavier than both the cassava and sorghum crops.

Perhaps the primary school teacher's most important task is to develop thinking and problem-solving skills. These skills are needed by all primary leavers--both by those who continue their education and by those for whom primary schooling is terminal. It is not certain to what extent thinking skills transfer readily from the situations in which they have been developed to the new situations in which they are needed. This is a topic on which more research is required. The CPE setters therefore decided to test thinking skills in as wide a range of contexts as possible. Thinking problems involving crop yields (such as no. 78) and crop rotation (no. 74) are likely to be relevant to the circumstances of many primary leavers in rural areas, whereas the problems involving acids and bases (no. 61) and temperature measurements (nos. 81 and 82) may be more appropriate to the cognitive needs of secondary school recruits.

Introduction of an Information-feedback System

The second main instrument of reform was to use the CPE not merely as a means for certifying pupils' achievements and selecting them for entry to secondary school but also as a source of information about the strengths and weaknesses of teaching and learning in the schools and, hence, as a means for improving the quality of basic education. An enormous amount of useful information can be generated as a by-product of examination processing at low cost.

The feedback information provided from CPE analysis was of two main types:

Incentive information consisted mainly of mean performance statistics for each school within a district and for each district within the country as a whole. The school performance lists enabled district support teams (supervisors, advisers and education officers) to identify the weakest schools within their jurisdiction which needed special attention; while the district lists enabled them to compare the overall performance of their own schools with the performance of schools in other districts.

Guidance information is more important than incentive information. It was based mainly on the analysis of performance in individual questions. The main guidance feedback document was the CPE Newsletter, which was distributed annually to all schools, district support teams, teacher educators, curriculum developers and other professional educators. The newsletter had two main purposes: first, to explain to teachers the changes taking place in the examination, both in the content of the questions and in the questions and in the cognitive skills being tested; and second, to identify key topics and skills which were causing pupils particular difficulties and to suggest to teachers ways in which they might help pupils to develop the necessary competencies more successfully. An attempt was made to write the newsletter in a simple, conversational style, avoiding the use of technical terms and complicated statistics.

An extended extract from the mathematics chapter of the 1980 newsletter illustrates the guidance feedback system. It had been found that pupils tended to perform better in the questions testing formal mathematics than in questions testing the application of number skills to everyday situations. Almost certainly, teachers had been devoting too much time to the more theoretical aspects of the curriculum and too little time to the practical aspects, probably because they saw the theoretical topics as being more important for the ablest pupils, who would gain places at secondary school. It was therefore decided to devote the entire chapter (25 pages in all) to a discussion of how teachers could develop in their pupils the skills needed to tackle number application problems more successfully. The extract gives the first four pages.

Extract from mathematics section of CPE Newsletter, 1980

Average marks in the 1979 CPE Mathematics paper were considerably lower than they were in 1978.

	Average marks 1979	Average marks 1978
Rural schools	43.7%	51.3%
Nairobi schools (excluding high-cost and private schools)	42.8%	51.7%

The average mark dropped by 7.6% in the rural schools and by 8.9% in Nairobi schools between 1978 and 1979. (High-cost and private schools are not included in the Nairobi sample). It is interesting to see that the rural schools overtook the Nairobi schools in 1979: their average mark was nearly 1% higher. In 1978, Nairobi schools performed a little better than rural schools.

Type of question	Number of questions	Average mark (rural candidates)
Arithmetic		
(a) Mechanical arithmetic	5	43.2%
(b) Applied number problems	21	36.9%
(c) Estimation of metric quantities	2	78.8%
Geometry	11	50.9%
Algebra	2	37.4%
Graphs	3	43.7%
"Modern" topics	6	45.3%
TOTAL	50	43.7%

The above table shows the average mark obtained by rural candidates in each of the different types of question included in the 1979 paper. You can see that blame for the poor performance cannot be placed on the "modern" questions. The 1979 paper contained only six questions testing topics which could not have been examined before the new KPM syllabus was introduced (for example, number bases, tessellation, and transformation geometry), and in these "modern" questions the average mark was in fact higher than it was in the 44 "traditional" questions.

It is clear from the table that the questions the candidates found most difficult were the applied number problems. The average mark in these questions was only 36.9%. The next-lowest performance was in the algebra questions (37.4%). Further, there were as many as 21 applied number problems in the paper, which is far more than for any other type. Thus the questions to which the paper gave most emphasis were answered least well.

24. A man working in a coffee factory is paid sh 2 per hour for the first 40 hours worked each week. For any extra hours worked that week, he is paid sh 4 per hour. In one week the man works for 64 hours. How much should he be paid for that week?

A	B	C	D
sh 128	sh 256	sh 284	sh 176

Question no. 24 is an example of an applied number problem. As you can see, the candidate is expected to apply his knowledge of the basic mathematical operations (addition, subtraction, multiplication, division) to solve a practical problem. He is not told which operations he should use; instead, he must work out from the given information which operations are needed. He must also decide to which numbers these operations should be applied, and in what order.

We can compare Question no. 24 with no. 9

9. What is 2·4 × 0·16?

A	B	C	D
38·4	3·84	0·384	384

This question is very different. The candidate does not have to take any of the decisions which are needed in no. 24. He is told which operation to use (multiplication) and which numbers to apply it to (2.4 and 0.16). Provided he can carry out the necessary calculation accurately, he is sure of getting the correct answer. Question no. 9 involves only mechanical arithmetic, whereas no.24 involves problem solving.

In the real-life use of mathematics, questions similar to no. 24 are far more common than questions similar to no. 9. Usually, we do not have anyone to tell us what operations to carry out to solve the problem; we must decide this for ourselves before starting the calculations. For this reason a high proportion of the questions in the CPE mathematics paper are applied number problems. It is disturbing that candidates find them so difficult. You should make sure that your pupils have developed the skills needed to solve these application problems before they sit the examination.

Steps in solving applied number problems

There are three main steps in solving applied number problems:

1. Understanding the information given by reading the question carefully
2. Deciding what information must be calculated, and in what order, to get the answer
3. Carrying out the necessary calculations in the correct order.

For Question no. 24, these steps can be set out as follows:

1. Information given
    - (a) The man works 64 hours in a week
    - (b) He is paid sh.2 per hour for the first 40 hours
    - (c) He is paid sh.4 per hour for hours above 40.
2. Information needed
    - (a) How many hours above 40 the man works
    - (b) How much he is paid for the first 40 hours
    - (c) How much he is paid for the hours above 40
    - (d) His total pay for the week

CPE Newsletter extract (continued)

3. Calculations

(a) The man works 64 hours in the week

Therefore the hours he works above 40 are (64-40) = 24

Therefore the man earns

(b) for the first 40 hours: 40 hrs x sh.2 = sh.80

(c) for the 24 hours above 40: 24 hrs x sh.4 = sh.96

(d) Therefore his total pay is sh.176

You can see that the calculations needed in step 3 to solve this problem are straightforward: one easy subtraction, two easy multiplications, and one addition. If the CPE candidates had simply been told to carry out these calculations, it seems certain that they would have found the question quite easy. But in fact only 23.6% answered correctly, and the question was one of the most difficult in the whole paper. From the wrong answers chosen, it is clear that the main reason the question was difficult was that candidates failed to complete steps 1 and 2 correctly: either they failed to understand the information given, or they failed to decide which calculations were needed. The percentages choosing each of the four answers were as follows:

	A sh 128	B sh 256	C sh 284	D sh 176
Percentage of candidates giving each answer	26.9%	32.9%	15.6%	23.6%

The two most common wrong answers were sh 256 (B), given by 32.9% of candidates, and sh 128 (A), given by 26.9%. The first of these two groups of candidates simply multiplied the total number of hours worked (64) by one of the two given rates of pay (sh 4 per hour). The second group multiplied the total hours worked by the other rate of pay (sh 2 per hour). Neither group took notice of the information given in the first and second sentences, which makes it clear that the man is paid at two different rates; the lower rate for the first 40 hours, and the higher rate for any extra hours.

Pupils cannot learn to solve application problems by following a set of rules, because each problem must be tackled differently. The best way to develop the necessary skills is to work through a number of problems with your pupils following the three main steps we have just discussed. You should give special attention to the second step. Deciding which calculations are needed, and in which order they should be carried out, requires careful reasoning; and your pupils will only develop this skill if they get plenty of practice. Do not tell the pupils what to do, but rather get them to make suggestions. Make sure they understand the importance of carrying out the calculations in the correct order. If you were teaching Question no. 24, for example, you should discuss with your class why it is necessary to calculate how many hours above 40 the man works (2a) before calculating his pay for those hours (2c) or his total pay (2d).

The development of numerical problem-solving skills is not a task for the CPE teacher alone, but should begin in the lower primary school. Every time a new arithmetical operation is introduced pupils should work through a number of story-problems involving the new operation. The next stage is to introduce problems in which pupils must decide for themselves which operation is needed. Finally pupils can begin to tackle more complicated problems of the kind we have just discussed, in which several operations must be performed, in the correct order, to reach the answer.

Real-life number problems are nearly always application problems, and they usually involve more than one operation. A farmer, for example, may want to work out how many bags of fertiliser he should buy for his coffee trees, or how many sheets of mabati he needs to roof his new house, or how much money he should receive from the dairy cooperative society for his milk at the end of the month. Each of these problems requires at least two operations for solution. To get the correct answer, the farmer must of course be able to add, subtract, multiply and divide accurately. But skill in arithmetical calculation is not enough by itself. To tackle the problems successfully, the farmer must also be able to decide which operations are needed, to what numbers they should be applied, and in what order.

You can see that skill in solving applied number problems is very relevant to the needs of CPE candidates who do not go to secondary school, and who are therefore likely to spend their lives working as farmers, manual workers or businessmen. The same skill is also essential at secondary school, both in mathematics and in the science subjects. For these reasons, application problems are now being given increasing emphasis in CPE.

It is disturbing to find that in the 1979 CPE, more than 70% of candidates could identify the solution set of the statement $1 < x < 5$ plotted on a number line (question no. 39): but less than one-quarter of the same candidates could correctly solve the applied number problem we have just discussed (no. 24). As we have seen, the average mark in the 21 applied number problems of all types included in the 1979 examination was only 36.9%. Results such as these suggest strongly that teachers are giving too much attention to the more theoretical parts of the KPM course and too little time to developing practical number skills.

Because they are so important, we shall discuss other applied number problems which gave 1979 CPE candidates difficulties. We can divide them into several types:

1. Money problems
2. Problems involving perimeter and area
3. Problems involving metric measurement
4. Time problems

An explanation of the school classification system used at the beginning of the extract is necessary. Primary schools in Kenya can be divided into five main types: low-cost schools, medium-cost schools, high-cost schools, assisted schools, and private schools. About 98 percent of schools in rural areas and 70 percent of schools in urban areas are low-cost schools. Pupils are charged no fees, although they are usually expected to contribute in a variety of ways to equipment and building costs. Medium-cost, high-cost, and assisted schools, by contrast, charge fees, ranging up to approximately $US50 per annum. Pupils in these schools tend to come from more privileged families. In a study of four high-cost schools in Nairobi, for example, it was found that 42 percent of the fathers of grade 7 pupils, and 18 percent of the mothers, were university graduates. Finally, a small but growing number of private schools charge fees ranging up to $US200 per annum and provide education of very high quality.

When low-cost schools alone are compared, no difference whatsoever emerges in the average mathematics performance of rural and urban schools. The same is true in all other CPE subjects, with the exception of English, in which urban schools have a 5-10 percent performance advantage.

A final point about the information feedback system is very important. The system could not be introduced until reforms in the types of questions being asked in the examination were well under way. If incentive feedback had been introduced while the examination still tested mainly rote recall, the effect would have been to encourage teachers to devote even more time to examination drilling. The introduction of feedback had to wait until it was clear to everyone that changes were taking place and that new preparation methods were needed. The first guidance information was sent to the schools in 1976; the first incentive information in 1977.

Some Effects of the Reform Program

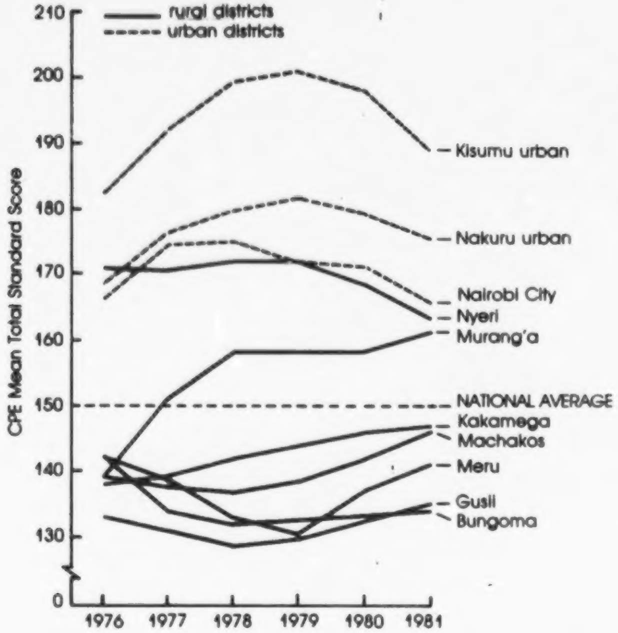
It is not possible to discuss here all of the effects which the reform program had on performance patterns. We shall therefore focus on only two.

The Effects of the Information-feedback System on Quality Differences Among Districts

It had been hoped that the introduction of our information feedback system would result in a reduction of the performance differences among districts. The weaker districts would have more to learn from the guidance feedback than the more successful districts, and so would be able to narrow the performance gap.

The actual effects were more complex. Figure 10.1. plots the mean total standard scores gained between 1976 and 1981 by the four districts which were most successful in 1976 and the six districts which were least successful. It will be remembered that standard scores give a measure of relative rather than absolute performance: the mean total standard score for all candidates is always one hundred fifty points each year. Kenya has too many districts to plot results from all of them on a single graph. But the overall trends can be summarized conveniently in terms of changes in the scatter of the distribution of district mean total standard scores from year to year. The measure of scatter which has been used is the standard deviation (SD). These are set out in Table 10.1. for the years 1976 to 1981. The first line shows the SDs for all forty-three districts in Kenya; the second line, the SDs for the thirty-nine rural districts only.

Figure 10.1. Mean Total Standard Scores for the CPE Examination, 1976–81



Note: This figure plots the mean total scores gained between 1976 and 1981 by the four districts which were most successful in 1976 and the six districts which were least successful in that year.

World Bank—40245:5

Table 10.1. Standard Deviation of Kenyan District Mean Total Standard Scores, 1976-81

	Standard Deviations					
	1976	1977	1978	1979	1980	1981
All Districts	9.84	11.57	12.64	13.09	11.79	10.15
Rural Districts Only	8.32	9.04	9.35	9.36	8.14	7.14

It will be seen from both the graph and the table that between 1976 and 1979 the differences in performance among the districts increased rather than decreased. The standard deviation of the means for all districts rose from 9.84 to 13.09; for rural districts only, from 8.32 to 9.36. The four districts which were most successful in 1976 increased their comparative advantage by 1979. Three of these districts were in urban areas. Conversely, four of the six bottom districts fell even further behind.

Between 1979 and 1981, however, these trends were strikingly reversed. The bottom six districts, without exception, showed substantial improvement, and the four most successful districts lost some of their lead. The standard deviation of the mean for rural districts dropped from 9.36 to 7.14-- lower than it had been in 1976.

These results suggest strongly that in the initial years, 1977 to 1979, the most successful districts, which had the strongest professional support teams, were able to benefit most from the new information available to them. In these districts, teachers' advisers and supervisors ran workshops and seminars to explain to teachers how they could equip their pupils to meet the new cognitive demands that the examination was now making. They singled out schools with particularly poor results for special attention. One such school, in a rural district with about two hundred fifty schools, improved its performance from being among the bottom 20 percent of the schools in the district in 1976, to being fourth from the top only two years later. Many other examples could be given of the capacity of primary schools in both rural and urban areas to improve the quality of their performance quite suddenly.

In the least successful districts, by contrast, the response to the introduction of the information feedback system was delayed. These tended to be the larger districts, either in terms of geographical area or pupil numbers, or sometimes both. Consequently, the professional and administrative infrastructure supporting the schools tended to be weak. Performance levels in these districts probably did not decline in absolute terms between 1976 and 1979, but because other districts were improving, standard scores dropped.

After 1979, however, these lagging districts began to respond. One major factor was increased public awareness. The district performance lists were made generally available in 1978, and low-scoring districts immediately came under considerable scrutiny. Almost certainly, it was this public concern which was mainly responsible for the upturn in the performance of the weakest districts between 1979 and 1981.

But the most striking trend has not yet been mentioned. In 1976 Muranga was among the weakest districts in the country; only five years later, it was among the strongest, offering a challenge to Nyeri, the top rural district, and even to Nairobi city.

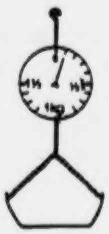
Muranga is a rural district, with about two hundred primary schools, none of which are fee-charging. (In Nairobi, as we have seen, about 30 percent of schools are fee-charging.) In 1977 the district support system was largely restructured. A number of new teachers' advisers, supervisors and education officers were appointed, and an intensive program of in-service courses and school visits arranged. The results which were achieved within such a brief period are impressive evidence of the scope which exists--in Kenya and almost certainly in other developing countries--for improving the quality of basic education through strengthening professional and administrative support systems.

CPE Science Items about Weighing, 1979 – 1981

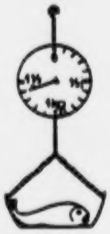
1979 62. Said goes to the market to buy a fish. He notices that the pointer of the balance is set as shown in Diagram P. When the fish is placed in the pan, the pointer is as shown in Diagram Q. Which one of the following is the most likely weight of the fish?

A. 1 kg. 100 g.
B. 1 kg. 200 g.
C. 1 kg. 300 g.
D. 1 kg. 400 g.

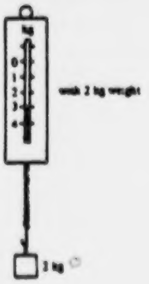
P

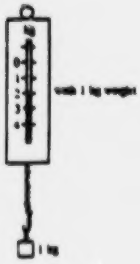


Q

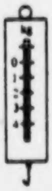


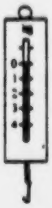
1980 67. Kamau thought that a spring balance was not showing the correct weight. The following diagrams show the positions of the pointer when he placed first a 2 kg. weight and then a 1 kg. weight on this balance.

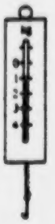


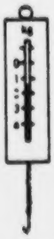


Which one of the following shows the MOST LIKELY position of the pointer when there was NO weight on the balance?


A

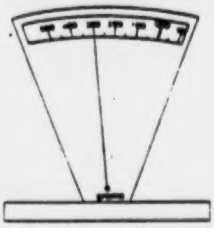

B


C

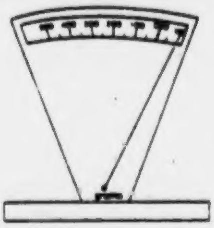

D

1981 55. Njoki weighed a small parcel before taking it to the Post Office. The diagrams below show the positions of the pointer on the balance when her parcel was on the balance and when it was removed.

With parcel on Balance



Parcel removed from Balance



What was the MOST LIKELY weight of Njoki's parcel?

A. 100 grams.
B. 200 grams.
C. 300 grams.
D. 400 grams.

The Effects of the Program on the Development of Specific Cognitive Skills

In the previous section we discussed the effects of the program on overall performance. Our dependent variable was the mean standard score for the whole examination.

We shall now take a much narrower perspective, and focus on the development of competence in a specific set of cognitive skills. Our dependent variable will be the percentage of pupils correctly answering the succession of three questions about weighing, which appeared in the 1979, 1980, and 1981 science papers.

The proportion of candidates giving each of the four possible answers to the 1979 question (Said and his fish) in rural low-cost and in Nairobi schools was as follows:

	A	B	C*	D
Rural low-cost schools	26.1%	14.6%	18.5%	40.4%
Nairobi low-cost schools	22.4%	17.2%	20.1%	40.3%
Nairobi high-cost schools	6.2%	7.0%	46.8%	39.9%
Nairobi private schools	3.6%	4.1%	54.8%	37.4%

*Correct answer

Clearly, the skills measured by this question are highly relevant to the needs of all pupils leaving primary school. Whether they continue with their education at secondary school or whether they begin work as farmers, craftsmen or traders, the ability to check that scales and balances have been set accurately will be very important. It was therefore disturbing to find that in low-cost schools, both in Nairobi and in rural areas, the question was very difficult. Only about 20 percent of pupils answered correctly, which is, of course, below the proportion we would have expected if all candidates had made no attempt to work out the answer, but simply guessed. In the fee-charging (high-cost and private) Nairobi schools, by contrast, the proportion answering correctly was about 50 percent. The performance gap between the two sets of schools was much wider than the gap for the science paper as a whole.

To answer this question successfully, candidates had to work through a sequence of three cognitive steps. First, they had to read the weighing scale accurately. Second, they had to transform the fractional notation used on the sclae to the decimal notation used in the alternative answers. Finally, they had to adjust the weight for the incorrect setting of the scale.

Analysis of the answer patterns suggests that the third step in this sequence was the most difficult for pupils in all types of school. The most common wrong answer in all samples was 1kg 400gm (D), which was, of course, the weight shown in the second diagram, with the fish in the pan. Clearly, a high proportion of pupils in all types of school did not know that if scales are not set to zero before weighing starts, they will show an inaccurate weight. In the fee-charging schools, most pupils worked through the first two steps successfully and were consequently able to reject 1kg 100gm (A) and 1kg 200gm (B); in the low-cost schools, substantial numbers of pupils chose these answers.

Given the relevance of the cognitive skills tested by this question, it was decided that similar questions should be included in subsequent CPE papers, but that in the meantime guidance should be provided to the schools and the school support teams through the CPE newsletters. The 1980 CPE newsletter, for example, discussed the weighing question from the 1979 science paper.

It will be seen that although the questions asked in 1979, 1980, and 1981 all involved similar situations and similar cognitive skills, the sequence of steps which the candidate had to carry out in order to solve the problem was quite different. In 1979 the candidate was given the two readings from the scale and asked to work out the correct weight, whereas in 1980 he was given the correct weight and the final reading and was asked to work out the initial reading. In 1981 the question reverted to the 1979 structure (given the two readings, work out the correct weight) but with an important difference: an extra step was added to the sequence needed to solve the problem. Once the candidate had recognized that the reading with the parcel in the scale was inaccurate, he had to decide whether this reading was too low or too high. In 1979 the "too low" option (1kg 500gm) was deliberately omitted, to make the question easier.

There were also other differences. First, the type of scale used was much less familiar than those used in the 1979 and 1980 questions (it is in fact the scale used in Kenyan post offices); second, the scale had to be read from right to left instead of from left to right, as is more common; and finally, the weight shown by the scale was too light, instead of too heavy, as it had been in 1979 and 1980. Any candidate who had been drilled by his teacher with the correct method for answering the 1979 question, but who had not developed an understanding of the general principles involved, would almost certainly have answered the 1981 question incorrectly.

Figure 10.2. graphs the performance of pupils in this sequence of questions, from 1979 to 1981, in Nairobi high-cost and private schools, in the six rural districts with the best overall results in the 1981 examination, and the six rural districts with the poorest results.

It will be seen that in 1979, all rural districts found the weighing question extremely difficult. In the most successful district, only about 22 percent answered the question correctly. Over the next two years, however, the top rural districts started to master the skills involved in problems of this type very rapidly. By 1981 the two most successful rural districts, one with about two hundred fifty schools and the other with about one hundred fifty schools, were rapidly closing the performance gap on the Nairobi fee-charging schools, which as we have seen, recruit their pupils mainly from privileged families and provide education of a superior quality.

In the weakest rural districts, by contrast, very little change in performance was demonstrated. By 1981, these districts were all still scoring below the chance-guessing level.

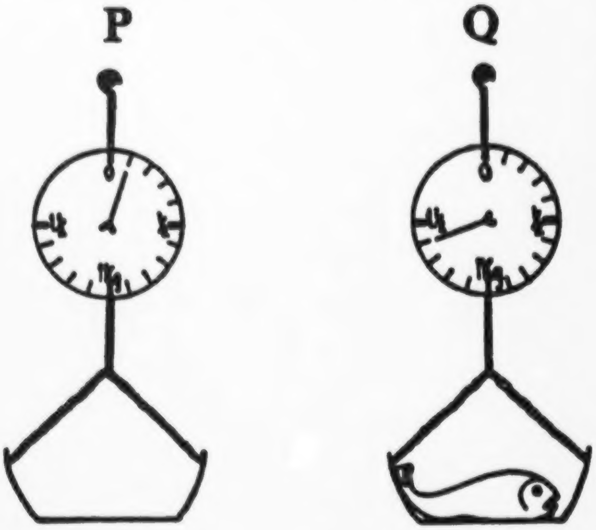
These results make it clear that the main reasons for the differences in performance between the Nairobi fee-charging schools and the rural schools are to be found more in the quality of education provided than in the home backgrounds of the children. In all rural districts in Kenya, a majority of the pupils are the sons and daughters of peasant farmers. As we have noted, one of the main differences between the most successful and the least successful rural districts in Kenya is in the strength of their school support systems. In the most successful districts, schools receive regular guidance from advisers, supervisors, and education officers, whereas in the less successful districts, the support teams are much less effective.

Extract from 1980 CPE Newsletter, discussing weighing questions from the 1979 science paper.

One of the questions in the science section of the paper which tested a terminally-relevant skill deserves special mention because it was answered so poorly:

62. Said goes to the market to buy a fish. He notices that the pointer of the balance is set as shown in Diagram P. When the fish is placed in the pan, the pointer is as shown in Diagram Q. Which one of the following is the most likely weight of the fish?

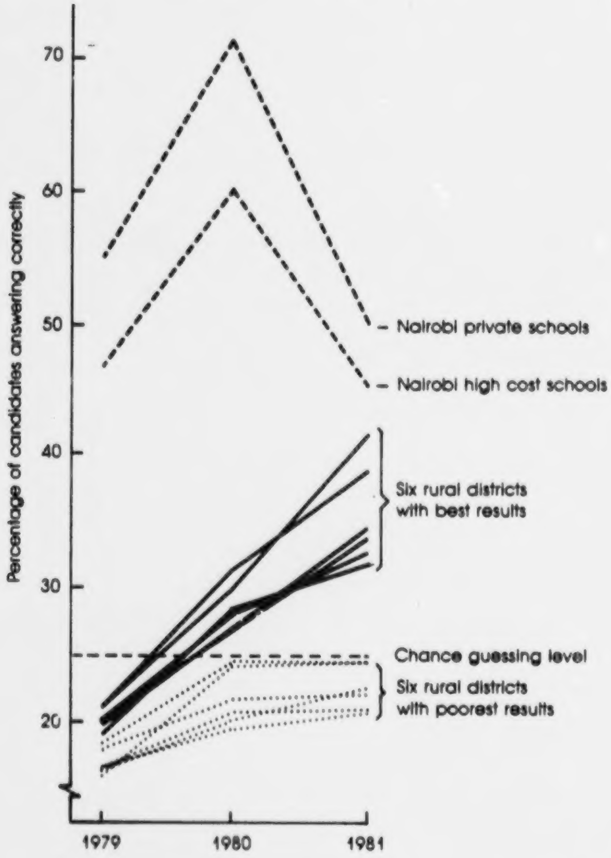
A. 1 kg. 100 g.
B. 1 kg. 200 g
C. 1 kg. 300 g.
D. 1 kg. 400 g.



Only 18.5% of rural candidates, and 20.1% of Nairobi candidates (excluding those in high-cost and private schools) correctly worked out that the most likely weight of the fish was 1 kg 300 g. More than 40% of candidates in both samples chose 1 kg 400 g. This, of course, is the reading shown in the second diagram (Q), with the fish in the pan. These candidates ignored the information given in the first diagram (P), which shows that before the fish was placed in the pan the pointer was set to 100 grams. Thus the fish probably weighs 100 grams less than the reading shown in Q. If Said made the same mistake as these candidates, he probably paid too much for his fish!

When a customer buys sugar, meat or fish from a shop, or when a farmer sells his pyrethrum, coffee or maize, he needs to be able to check that the weighing is done accurately. During geography and science field trips, pupils should become familiar with balances and scales of as many different types as possible. Pupils can observe weighing being carried out in places such as coffee factories, post offices and tea 'centres' (places where tea farmers sell their green tea leaves) as well as shops and markets.

Figure 10.2. CPE Performance in Weighing Questions, 1979–81



Note: This figure plots the performance of the six rural districts with the best overall results in the 1981 examination and the six rural districts with the poorest results.

World Bank—40245:9

Clearly, a reform program for rural education based on the rengthening of school support teams, and using the analysis of examination scripts as a source of insight into the reasons for pupils' cognitive difficulties, could have a major impact in improving the overall quality of education and in reducing quality differences among schools.

IV

SUMMARY

11

IMPROVING UNIVERSITY SELECTION, EDUCATIONAL RESEARCH, AND EDUCATIONAL MANAGEMENT IN DEVELOPING COUNTRIES: THE ROLE OF EXAMINATIONS AND STANDARDIZED TESTING

Stephen P. Heyneman

Common Background to Testing

Why is it that widely diverse countries have such a deep interest in educational testing? One reason is the high premium both rich and poor countries place on the ability of their educational systems to accomplish national goals. A systematic and reliable mechanism for monitoring achievement performance and for identifying individuals with aptitude and ability is indispensable to achieving these goals.

Political Consequences of Educational Failure

The idea that economic productivity is affected by the level and nature of educational investment (human capital theory) is no academic fiction. Regardless of political party and independent of ideological view, national leaders quickly become alarmed at reports of a decline in educational performance. They believe national interests to be at stake. They believe that schools teach the necessary technical skills to fuel future industrial and agricultural innovation and progress and that schools teach the behavior patterns necessary for survival of policy-- national history, language, social obligations. In certain specific ways the educational system is held accountable for national economic potential and domestic social behavior. Consequently, the stake in education is perceived to be very high. If the behavior of young people is not acceptable, if their intellectual performance is lackluster, if the portion of young people devoting themselves to more difficult fields of mathematics and science declines, these signs are interpreted as failures of the educational system. Because the stakes are high, the political consequences of failure are also high. They can precipitate a national crisis. The crisis itself normally generates public investigations, reforms, and shifts of resources. The political consequences are a reason why widely diverse countries maintain such a high interest in techniques of educational testing, for the results of the tests, to some extent, determine the degree to which the system is meeting its expectations.

Social Effects of Testing

The social effects of testing are felt in all countries. "Social" in this instance is used to differentiate between the individual's psychological effects (anxiety, for example) and the structural effects on the educational system. Are educational systems affected by the use of tests? The answer is "yes" and in two specific ways.

210

First, testing produces tangible criteria by which to judge education quality. Systems of education without selection examinations, or in which testing for assessment purposes is ad hoc, do not have a tangible means of judging quality. This is true of many educational systems in Latin America, where judgments on quality are limited to statistics on dropout, progression, and enrollment rates. Systems of education which do utilize testing functions are subject to evaluations and public pressures of a different sort.

The pressures felt by educational systems which use standardized tests are broadly popular since all sections of the population are concerned, and since, in spite of technicalities, scores can be broadly interpreted. Even when results appear to affect certain sub-groups detrimentally--blacks in the U.S., working class youth in the United Kingdom--the testing function itself is powerful. It draws attention to a single, widely understood indicator. It holds the school system accountable for results, and it fosters a continuing open forum on the school system's ability to provide high-quality education. Debates do not occur with such vigor where there is no standardized testing.

Second, testing provides a universalistic criterion for choosing those who qualify for further training and for later societal leadership. To be sure there are those who argue that testing is tyranny, that it creates unnecessary anxieties for the individual, and that it is inherently biased against certain sub-groups of the population. Despite these persistent charges the fact is that no criterion--other than random lottery--has been found to be quite as fair. No criterion has been found to be more valid in determining future success. When school systems have relied upon more particularistic criteria for selection to higher education--political loyalty, family alumni status, personal wealth, ethnicity, geography of birth--the effect has been more pernicious. Those left out feel (with justification) that the choice was made unfairly, as in the case of political loyalty; or on the basis of ascription rather than achievement, as in the case of ethnicity or wealth.

That the results of testing are not random, however, is a concern common across countries. Certain kinds of tests are better than others. Certain methods of setting questions are more fair or more reliable. No matter how professional, tests are inadequate for many kinds of judgments. These facts helped create the high degree of interest in testing technique and testing policy among the educational professionals in China attending this seminar.

In spite of the universalistic nature of standardized testing, political pressures attempt to divert the criterion from one strictly based on achievement. These pressures are common to all countries regardless of ideology or history. There are pressures to rectify past injustices to minorities, pressures to ensure fair geographical representation, pressures to recognize and reward abilities other than traditional academic ones. How should testing professionals respond to such pressures? Is there an ideal combination of selection criteria? Can results be less biased and so reduce political pressures? These are the questions of concern.

Testing Issues

In most Organization for Economic Cooperation and Development (OECD)-member countries, educational testing is a sizable industry. In the United States and Australia, this industry is very competitive and is characterized by large numbers of private firms. Their clients consist of the numerous public or private schools and school systems. However, a supply of private firms testing for assessment does not necessarily imply a similar supply testing for college selection. With sixteen thousand school districts in fifty states, the U.S. still relies for the most part on only one or two testing agencies for the design and administration of selection tests. By contrast, with primarily a unitary, centralized school system, Sweden goes to the opposite extreme in terms of university selection. In Sweden performance ratings are designed and calculated by each classroom teacher. Though not independent of centralized standards, nevertheless these performance ratings are independent of centralized control. The United Kingdom has more than a dozen examination boards, though the population being examined is only 10 percent of the United States, Australia has eight different selection examinations, a different one for each state. Japan has two sets of examinations. The first one is unitary throughout the nation. The second is diversified, with different formats, emphases, and subjects administered independently by each college and university on different dates, at different places, and at different time. In consequence Japan has hundreds of different selection examinations.

Such varied national systems would seem to suggest totally independent problems, but this is not the case. In reality many problems are common to all countries and pertain equally to all educational systems and testing enterprises. Some are unresolvable, that is, they are of such complexity and pose such deep dilemmas that no simple solution exists. Solutions can only be found through the study of certain local principles and the application of those principles to new environments -- in China or in other countries where reforms are being considered. A list and a discussion of these testing dilemmas follows. Yet not all principles emerging from this workshop are unresolvable dilemmas. Some in fact are generalizable. These are listed and discussed in a subsequent section entitled "general recommendations".

Aptitude Tests vs. Academic Achievement Tests

Considerable debate occurred over whether it is more useful to test for academic achievement or for academic aptitude.[1] Since the 1940s most American colleges and universities have been assessing student candidates on the basis of their scholastic aptitude. The Scholastic Aptitude Test (SAT) was used in Japan in the 1950s and currently is widely, though not uniformly, used by local testing authorities in Australia. The SAT has not been used in Sweden, Britain, or France.

Four principal drawbacks and two principal virtues of the SAT were discussed during the seminar. Though designed to reflect basic abilities, the first drawback is that the SAT is, in fact, subject to coaching. Disagreement existed over the degree of the coaching effect but not whether it exists. On the other hand, many proponents of academic achievement testing do not view coachability as a drawback but rather as a virtue. One reason why the SAT was rejected in Japan, for instance, was because it was not as coachable as an achievement test. The second drawback is the SAT's predictive ability. Japanese,

---

[1] Like other tests of aptitude, a test of scholastic aptitude is designed to utilize the general principles one is likely to have encountered in an academic curriculum but not the specific knowledge one is likely to have encountered in any one state, district, or classroom. In principle an aptitude test is broadly applicable across the sixteen thousand U.S. school districts, each with its independent curricular authority.

Australian, and American experience was uniform--university performance is more closely predicted by academic achievement than by academic aptitude.[2]

Thirdly, the SAT creates a distance between classroom and test. It limits the ability of teachers to prepare students and thereby lowers the feedback effect of past test results. When performance is poor in a mathematics examination based on the curriculum, strategies of amelioration are significantly more evident and more clear than if the poor performance had occurred on the mathematics section of an SAT.

The last drawback is the issue of diligence. Performance on an academic achievement test is in part a reflection of a student's diligence, his ability to study hard for long periods of time, his maturity of purpose. Diligence is one of the principal ingredients of university and later professional success. And many feel that diligence is one of the key ingredients missing among students in the American system of higher education.

But the SAT does have several powerful elements to recommend it. Because it is not criterion-referenced and is not based upon a single concept of subject excellence, it frees local school districts to experiment with creative curricular innovations. Designed and set centrally, an academic achievement test will dictate to a very large extent what is taught in the classroom. In sum, the limited feedback power of the SAT which some see as a drawback, others see as a virtue. Because pedagogy is less affected by test results, pedagogy can be more experimental.

Equity is also a principal virtue of the SAT. The SAT is more able to identify bright, low-income students from impoverished home environments and from impoverished schools. For example, though only 5 percent of the students in the United Kingdom attend private (fee-paying) schools, 45 percent of the incoming entrants to Cambridge University come from private schools. They have attended better schools and they have performed better on the academic achievement tests used for university entrance. But are they the best and the brightest among their age cohort? This question is especially serious in developing countries, and in China in particular, where the quality of education varies so markedly between city and province, wealthier suburb and impoverished commune. What an academic achievement test measures in this circumstance is very largely the opportunity to learn, and much less so the ability to learn. Any nation concerned about picking its future talent pool must consider the proven ability of a test such as the SAT to compensate for the differences in classroom facilities which preclude an equal opportunity to learn.

School-based Assessment vs. External Assessment

Substantial time at this seminar was devoted to discussion of assessments of the school-teacher grades and teacher-designed achievement tests. These discussions were not planned, since school-based assessments are not usually considered part of standardized testing, at least not in the usual sense. Rather the subject arose because of its importance.

---

[2] On the other hand, the Educational Testing Service (ETS) accurately pointed out that university performance was more closely predicted by score on the SAT in combination with academic (school-based) grades than by the SAT alone or by grades alone. Most American colleges and universities use a combination to make selection decisions.

School-based assessment is very popular among teachers. It gives them a sense of efficacy. It provides them with an important role in the university selection process. A school-based assessment is able to measure student characteristics from selection examinations and can therefore influence what is thought to be either necessary or superfluous with regard to examination content. For instance, the United States does not have essay questions as part of its SAT, but most American universities base their selection decisions on a combination of grades plus SAT scores. Since grades ideally should contain a measure of writing ability, many argue that the SAT need not duplicate what is already being tested through school assessment.

Similarly, in Sweden it is argued that school-based assessments are able to measure what standardized tests cannot--student character, diligence, the will to succeed, and most important, the improvement of these qualities over time. Teachers are likely to know if a student is working up to potential; they should be able to tell if someone who has had difficulties concentrating has finally found motivation. They should know, in essence, the quality of the student's knowledge. These factors all make up part of the school-based assessment. For these reasons such assessments are excellent predictors of first-year college performance and are growing in political popularity. School-based assessments have all but replaced nationwide selection examinations in Sweden. In the U.K. school-based assessments are provided by the Local Educational Authorities (LEA). The case of the U.S. has already been mentioned. In Australia and in Japan school-based assessments records are made available to universities for making selection decision. However, the weight placed on such information varies.

Testing experts feel ambivalent about school-based assessments. Such assessments are not standardized. Criteria vary from school-to-school and even from teacher-to-teacher. What may constitute an "A" grade, may be only a "B" grade elsewhere. In some schools equal weight is placed upon physical education, fine arts, and physics. In other schools only academic subjects figure into the grade-point average. Moreover, school-based assessments vary over time. The U.S. in the 1960s is the premier example. At that time student privileges and rights were in the ascendency. Curricular requirements were diminishing. Teachers tended to assess students according to less rigorous criteria than they had ten years earlier. Such grade inflation made the selection decision less accurate and more difficult.

On the other hand, there are examples--often from Sweden--of efforts to minimize these problems. Sweden has gone very far toward making school-based assessment the sole criterion of selection. Its system functions despite the problems acknowledged above because (a) the student population is very small and (b) efforts to minimize the variation in teacher-grading criteria have been elaborate. All student results are compared to national norms in each subject; criteria of subject assessments are regulated by means of national guidelines; and when extreme variance occurs, that is, when results from one teacher, one school, or one student to another stand out in some extraordinary fashion, then committees are appointed to investigate the source of the problem and to recommend adjustments. By these means school-based assessments have become the national methods of selecting university entrants.

For large countries, such as China, the use of school-based assessments in lieu of standardized examinations is problematic. Logistical problems become paramount. Issues of unexplained variation would create political difficulties. Means of enforcing standardization procedures, such as those used in Sweden, would be too cumbersome in such a large country and more suspectible to unprofessional manipulation. Countries with large student populations have no choice but to utilize standardized selection examinations.

However, school-based assessments should have a role in the selection decisions of any country. They do provide more information and more personal information. Chinese universities ought to have school-based assessments at their disposal. They would then be able to consider the diligence of specific pupils; or individual efforts to support community projects; or ability in sports, music, and art. All of these qualities make up the character of people whom universities may wish to select. Having school-based assessments would make whatever special emphasis they choose possible.

Multiple Choice vs. Non-multiple Choice

Multiple-choice formats have been in use in standardized testing since World War I and are dictated by virtually irrefutable necessities: they are easy to score and they can be given to large numbers of individuals quickly and cheaply. But seventy years after their first appearance, controversy remains. Many feel that the shortcomings of multiple-choice formats outweigh the benefits. The Cambridge University Examination Syndicate, for example, relies very heavily upon non-multiple-choice formats of examination. At both the "O Level" and "A Level" the syndicate administers essay-formated tests. Swedish classroom teachers, many Japanese universities, and some Australian states rely on the essay format, as do testing officials in France and Germany.[3]

In each national case, it is clear that teachers generally prefer non-multiple-choice test formats. Essay questions, for instance, are more typical of normal classroom discourse. They represent what occurs most naturally--a student-created response. Whether in school or work settings people are rarely faced with preprogrammed choices.

In each testing situation a hierarchy of skills exists which tests wish to measure. The simplest are knowledge and vocabulary-awareness skills. In order of complexity follow comprehension, application, analysis, synthesis, and evaluation skills. The consensus is that multiple-choice formats are more efficient at measuring skills at lower levels of this hierarchy. Proponents of non-multiple-choice formats claim that they are better (though not necessarily more efficient) at measuring skills of synthesis. Synthesis skills consist of those which one utilizes to combine many pieces of knowledge, facts, and principles, with the end result being a single product. Proponents claim that an essay test is a better indicator of skills of synthesis.

According to proponents of non-multiple-choice formats, they are amenable to the same types of reliability and validity reference points as are multiple-choice formats. Non-multiple-choice formats can be standardized. They can be objective. What is required is a carefully constructed system of professional test graders -- as exists in the United Kingdom -- with strong quality control, clear criteria of excellence, and effective internal procedures.

All test experts agree that good multiple-choice test items are difficult to construct. They have to be clearly written; their choices have to be plausible and comprehensive. Test designers need to know the principle on which the correct answer is based and also the variant principles on which all conceivable incorrect answers are based. This requires considerably more research capacity than is normally available at the school level. Consequently, good multiple-choice tests are rare, while poor ones are common. Each of these arguments made in Beijing points toward the use of non-multiple-choice formats.

---

[3] Essays are only one form of non-multiple-choice test formats. Others are fill-in-the-blank, the short answer, and design projects. At the "A Level" in highly selective fields such as engineering, the Cambridge University Syndicate administers tests of originality, that is, a design project worked on for six months or more by each individual candidate.

Proponents of multiple-choice formats, however, are quick to argue their strengths. When carefully designed, multiple choice is useful even for testing higher-order skills of analysis, synthesis, and evaluation. Mr. John Keeves, for instance, discussed specific examples of these higher-order multiple-choice questions designed for the IEA science tests. Though difficult to accomplish, multiple-choice formats can be designed to assess skill involving creativity.[4]

Multiple-choice formats are amenable to scientific techniques of pretesting on an item-by-item basis. The SAT contains 20 percent "dummy questions", questions placed in the test solely for purposes of experimentation. Thus, when a test question is ready to appear in final form as an ability measure, test designers already have a fairly exact understanding of how that test item will perform even before it is actually administered. They know ahead of time, for instance, the degree of test bias -- against southerners, or northerners, against children who went to Catholic schools, against non-English speakers, or against females. All of these can be predicted ahead of time with carefully applied multiple-choice test formats. This does not mean that the reliability of non-multiple-choice formats is questionable. What it does mean is that if they are, the problem will more likely be discovered after test administration rather than before.

Among the five OECD countries no examination agency relied exclusively upon a single test format. The Cambridge University Examination Syndicate is known principally for its non-multiple-choice formats, but it also uses multiple-choice formats. Similarly, the Educational Testing Service (ETS), while known for its use of multiple-choice tests, annually administers thousands of written examinations as well.

These advantages and disadvantages of different test formats often cannot be separated from the experiences of test application. Thus, the method of multiple-choice, aptitude-test application specifically used in the United States has one advantage over the method of achievement, essay-test application specifically used in the United Kingdom. This is the ability to generalize from year to year and from subject to subject. Because examinations in Britain are designed by specialists against the criterion of subject matter excellence, the proportion of students who receive a grade of "one", or "two" or "three" varies from one year to the next. Variation may occur either because students perform better from year to year or, alternatively, because the difficulty of the examination may vary from year to year. By basing test results on normalized scores, the American system is not plagued by variations in the distribution of those scores. This allows comparisons across subject matter and from one year to the next.

It appears, however, that the most important factor in deciding between multiple-choice or non-multiple-choice formats is not test theory but test economies. Non-multiple-choice formats are anywhere between two to five times more expensive to design and grade. The larger the tested population, the larger the cost differences become. Though economies of scale exist in both formats, the marginal cost of adding a tested pupil in the U.S. is only a few pennies once the test is designed; the marginal cost in Britain is substantially higher due to the very high cost of test marking. These economic and logistical facts emerged again and again in this seminar. The ETS annually examines many times the number of students examined by the Cambridge University Examinations Syndicate. An essay format for every American university candidate would place the cost of the examination significantly above what it is currently and perhaps out of reach of the average family income. These costs would have to be subsidized by the government. If not, children of the working class, who have fewer economic resources, would tend to be excluded from university opportunity not only because of university tuitions but also because of the cost of university entrance examinations.

---

[4] If creativity is defined as novel ways of looking at a problem, multiple-choice test formats can be used. Responses can be thought through ahead of time by the test designer. If creativity is defined as unique ways of looking at a problem, multiple-choice formats would be inappropriate; in fact, impossible since a criterion of excellence is something not yet invented.

China currently has a college entrance population at the level of the U.S., about 1.5 million. While this represents 40 percent of the American eighteen-year-old cohort, it represents less than 1 percent of the Chinese eighteen-year-old cohort. The proportion of the population going to universities in China is expected to rise, to 3 percent by 1990 and perhaps to 10 percent by the year 2000. This would entail a doubling, a tripling, even a quadrupling of the Chinese examination candidates over the next fifteen years. The force of these costs and the logistical management of the test process will require China to move gradually from an examination system based entirely upon non-multiple-choice test items, scored by hand, to an examination based principally upon multiple-choice items, scored mechanically.

Test Questions: Should They Be Public or Private?

OECD countries have very different traditions on the question of test privacy, and these different traditions have had pronounced effects on the education system at large. In those countries where test items are made public, in Japan and the U.K. for example, the effect is to generate "examination students," that is, students who spend a great deal of time studying portions of subject likely to be examined and methods of response likely to elicit a good score. It is alleged that this intensive examination preparation can detract from more creative study of subject matter and can generate a false standard in the utility of knowledge. Knowledge less likely to be examined is considered less useful. Publication of previous examinations, moreover, makes it necessary to rewrite new tests almost yearly. This requires much, perhaps unnecessary, research. It also raises the chance of test bias due to hasty design. According to some, this is too costly.

The SAT offers one example of the effects of test privacy. Test booklets are numbered and are returned after each test application. Then they are destroyed. This allows test items to be used again with little danger that students having taken the test previously will be at a significant advantage, and it dramatically lowers the cost of examination design.

Keeping tests private entails two problems. If applied to academic achievement tests the feedback mechanism will not be as effective. Classroom teaching is more effective when actual test examples are used. Moreover, political problems are associated with test privacy for both achievement and aptitude tests. Privacy places distance between the testing agency and the general public. The public is expected to accept that the test was a good test on the basis of faith or, alternatively, on the basis of statistical information that few can interpret -- reliability coefficients, Kuder-Richardson indices, factor loadings. Much simpler and much more direct is for the public to be able to read each question and to debate each question's virtues and drawbacks.

In developing countries faced with severe social and political volatility, public trust in institutions assessing further educational opportunity is essential. Thus, despite the cost savings which can be realized from test privacy, the wise choice in developing countries will be to make tests open and public each year and to readily encourage public debate on test content.

Admissions Decisions: Who Should Make Them?

In all countries university education is expensive and student places are scarce. Hence, it is easy to understand why governments, especially in developing countries, would wish to select who attends. After all, the nation's future is at stake, and what organization is more responsible for the nation's future than the government?

Governmental mechanisms for making admissions decisions vary. Most countries -- China and Tanzania, for example -- do not review each individual decision. Instead they establish such strict entry criteria that universities have little or no latitude for making exceptions.

Such is not the case in the U.K., the U.S., Japan, or Australia. In each of these OECD countries, governments have little or no role in the selection decision. To be sure, testing agencies, such as those in Japan, are often supported by governmental resources. Nevertheless, admissions criteria and decisions on individual admission are made independently by each university.

This was, in certain ways, revelatory to those in the audience from Chinese universities. They had long struggled with the problem of allocating students to one university or another irrespective of personal interests. They had long felt the need for their universities to develop individually, to specialize, for example, in engineering, literature, or economics. But such individual directions could not occur without control over who will be admitted as students.

Reports from the OECD countries present at the seminar seemed to confirm their interests. It was clear that universities in the U.K., the U.S., Sweden, Japan, and Australia are not identical in course offerings or prestige. It was also clear that open competition exists among them. Had Yale lost its edge in economics to Berkeley or the University of Chicago? Was the University of London more interesting now in the field of physics? Was business administration at Kyoto more advanced now than in Tokyo? These debates occurred among the presenters both formally and informally, with the effect of demonstrating the efficiency of giving universities more control over their own admissions. This appeared to be particularly important for rural, less prestigious universities. Without the ability to develop specializations it appeared that they would have no hope of becoming competitive with the major urban universities. Mention was made of literature at the University of Iowa, agriculture at the University of California at Davis, and petroleum engineering at the University of Oklahoma as instances of successful competition on the part of rural, less-well-known universities.

But could universities make selection decisions fairly? Would they not select on the basis of family privilege? Would they not discriminate unfairly against ethnic or religious minorities? Such problems were admitted in each of the OECD countries. But the role of the testing agency in OECD countries was clear: it was to test as fairly as possible, but to leave the selection decision to the university and, in cases of potential discrimination, to the university as modified by government.

Such issues were foremost on the minds of the Chinese participants who, at that moment, were engaged in a nationwide discussion on whether the government or universities should make the selection decision.[5] A substantial heterogeneity of views was expressed on the subject.

Technical specialists generally hold that testing is a question of technical professionalism and that the testing agency is best when it is independent of government and selection decisions. Reports from Japan, the United States, the United Kingdom, and Australia confirmed this intuitive view. University control of admissions assumes the independence of academic institutions in other areas as well -- the content of study, the criteria for academic excellence, and the direction of academic research. Who makes admissions decisions cannot easily be separated from other, wider implications. However, the fact that universities in all OECD countries are charged with these responsibilities made a significant impact on the audience in Beijing.

General Recommendations

All school systems share common managerial challenges, and as a result the educational leadership in each country is likely to learn a substantial amount by studying practices found to be successful elsewhere. Such comparative analyses are today the norm rather than the exception. Nevertheless, adoption of techniques and practices is unlikely to be productive if it is conducted without careful consideration. Traditions vary from country to country and so do the abilities and resources required to put reforms into practice. Adoption of new practices requires significant forethought among OECD countries; adoption of practices from OECD countries to developing countries requires even more.

This said, a short list of principles can still be drawn from these descriptions of the uses of examinations and standardized tests in Sweden, the U.S., the U.K., Japan, and Australia. These principles appear to be worthy of consideration in China and perhaps other countries as well. The recommendations might be organized into three categories: the requirements for the professionalization of testing, test administration, and test content.

Requirements For The Professionalization of Testing

Despite differences in size and resources, school systems in developing countries, without exception, require a professional capacity in the field of standardized testing and examinations. This has four prerequisites. The first is research capacity. Countries which set their own examinations must therefore develop them. Test development can be done in a haphazard manner, but the negative technical and political consequences are serious. Professional test development requires a systematic program of item design and experimentation; an on-going and permanent program of test-result evaluation; an active item bank; an extensive set of relationships in the higher education community as well as the teaching community in elementary and secondary education; and a minimum commitment to experimentation of new testing technologies and equipment.

Developing countries can rarely afford to establish a research capacity in the field of examinations as a separate entity from their research capacity for the more general purposes of standardized testing. Yet these two functions are usually kept separate. The first is situated in an independent testing bureau, the second in a research and planning bureau. Examination development is easier to professionalize if the examination bureau is autonomous from the main functions of government. Such autonomy need not arise at the sacrifice of an economy of scale in testing research and development. Computer and optical-scanning equipment need not be separate. There is no reason not to share (non-confidential) statistical data. There is no reason not to jointly plan equipment experiments or jointly participate in technical training. To professionalize testing, developing countries must develop research capacity; but given resource constraints, developing a research capacity in the field of examinations will not be feasible without coordinating its development with that of educational research in general.

---

[5] One month after this conference the government decided to shift the responsibility of admissions to the universities.

Second is training. The professionalization of testing in China, as well as in other developing countries, will require a regular program of training, both internal and external. The required skills are often broader than is commonly imagined. Clearly a professional testing capacity requires psychometrists and statisticians. It also requires survey research specialists, computer programmers, art and graphics designers, publishing specialists, production and distribution managers, cost accountants, and social scientists. Each has a specific function in the production process. Eventually a joint product is delivered at exactly the same point in time, with no production errors, to hundreds of thousands of individuals without fail several times per year. The logistics of professionally managing a system of testing requires training.

Third is equipment. No system of testing today is not dependent upon computing equipment. School-based assessments, even if graded manually by teachers, require a systematized memory. Essay tests, graded by hand, require statistical records. Fill-in-the-blank tests, creative design tests, and oral tests all require scores, statistical analyses, historical records, and systematic reporting. It is not necessarily true that the more sophisticated the equipment, the higher the level of testing professionalization. But it is true to say that professionalized testing requires electronic equipment.

Equipment, in turn, has three requirements: an adequate source of development capital, a careful plan of acquisition and utilization, and a carefully phased program to develop and to maintain the skills of technical maintenance. The degree to which the functions of examinations and standardized testing are localized will determine the complexity of the policy toward equipment acquisition. Each of China's twenty-nine provinces will require its own expertise and facility. The magnitude of students involved and the newly acquired provincial examination authority require professionalization on a province-to-province basis. In some countries, such as the United States, professionalized testing is found right down to the school-district level. This is because school districts design the curriculum and pay for it. The principle is generalizable though. The geographical unit in charge of curriculum design, school finance, or school selection will require a professional testing capacity, which, in turn, will require necessary equipment.

Last is the need for new ideas. This meeting demonstrated two factors to both Chinese participants and the official presenters from the OECD countries. It demonstrated how variable the traditions of testing and selection across countries are, and how much they are affected by each nation's political culture. Sweden's system is heavily influenced by its history of inequity in higher education participation and its current social democratic effort to overcome that history. The United Kingdom's system is heavily influenced by the prestige of its ancient university authorities and traditions of academic excellence. The United States' system was created as an outgrowth of its fierce curricular independence on the part of local educational authorities. Japan's system is affected not only by its traditions of university prestige, but by the cultural predilection for academic diligence narrowly defined around subject-matter authority. Australia's system is a hybrid of American and British traditions -- with a multiplicity of provincial differences and mixtures of academic achievement and scholastic aptitude. There is no single technically correct way to design a system of examinations. Despite the substantial and understandable desire on the part of the Chinese technicians to acquire new ideas on "how to do it", the inescapable fact remains that how OECD countries do it is heavily influenced by the different reasons for why they do it.

A second point made clear at this meeting was the degree of rapid change now occurring across the board. High-level debates on their respective examination systems ensued among politicians from China, Japan, the U.K., and the U.S. Major changes were anticipated--and in completely opposite directions. Japan is likely to diversify its examinations subjects; the U.S. is likely to concentrate them. The U.K. is likely to move away from nationally set examinations and toward school-based assessments. In Sweden and in Australia the trend is uncertain. What is certain is the degree of change.

Both observations have implications for China, and for developing countries in general. What is implied is that ideas for testing design and testing policy must be diverse across referent countries and must be kept up-to-date.

Test Administration and Management

The administration and management of selection examinations and the administration and management of standardized tests used for system-wide assessment are both important. Selection examinations play a key role in a nation's economic development and therefore are to be considered an important national resource which needs to be protected. The means by which this is accomplished in many OECD countries is establishing the examination agency as an administrative body autonomous from government control. The U.K., the U.S., Japan, and Australia allow the testing agencies to collect user fees for examinations. Income from these fees remains within the agencies themselves. The fees are small enough so that they do not inhibit the chances of attending university, yet they are large enough to allow the testing agency to build their own research capacity and to set their own standards of technical excellence. Professional standards cannot be maintained if administrative budgets are subject to the ebbs and flows of ministerial politics or national economic exigencies. Income derived from testing is inevitably secure since it is based upon a guaranteed demand. Independence of fiscal resources is a prerequisite for independence of professional standards.

None of the testing agencies from the OECD countries making presentations in Beijing assumed responsibility for the selection decision itself. Examining as a function is kept separate from the application of examination results. It was recognized that the university itself should make these decisions independently. It was normal for universities to place different weights on different portions of the examination and, more important, to place different weights upon the non-examined elements of a student's character. In sum, it was recognized that the selection decision should not be subject to a mechanistic system and that however they were constructed and whatever their content, examinations could only establish benchmarks of student accomplishments or potential. Moreover, examinations could not be expected to establish these benchmarks for all student characteristics.

Japan, the U.S., the U.K., Sweden, and Australia have very different examination systems, but none has a single examination agency with a monopoly. The National Center for University Entrance Examination in Japan serves as the sole source for examining the first stage of university entrance, but the second stage consists of a plethora of different examinations and competitive examination bodies. The ETS in the U.S., while widely utilized, is not the sole examining body. In each of these non-monopolistic cases, the testing agency is forced to maintain professional standards not only with the general public but in an actual competitive sense in relation with other testing agencies. This competition is healthy. It generates informed discussion, and it provides the consumer (for the most part schools and school districts) with alternative choices. Large countries such as China might be wise to develop a diversity of examination systems and agencies in some combination of local and national authority. Smaller countries may not be able to support such a diversity. In these instances consideration might be give to external institutions such as now exist on a regional basis in West Africa and the Caribbean. If a lesson from the examination experience of the U.S., Japan, the U.K., and Australia is to be gained, however, it is that competition works for the common benefit.

Uses of Standardized Tests for System-wide Assessment. The use of standardized tests in the management and administration of an education system goes much further than the selections for higher education. The political use of selection examinations is the best known of its other functions and was certainly of high interest to the Chinese in this seminar. Less visible, equally important functions were also discussed, and these next recommendations concern them.

(1) Measuring achievement over time. All school systems, in countries large and small, need to record their progress -- not just in the number of students educated, but in the knowledge the students have acquired. Such a record cannot be attained by studying the results of selection examinations, for only the best students who have finished secondary education take them. Such a record cannot be attained by studying the results of school-based assessments, for standards shift from year-to-year and vary from school-to-school. Measuring educational achievement over time requires a specially designed test, with "anchor" items to circumvent changes in curricula. It requires that it be given to a special sample of students at regular intervals.

In this seminar a presentation was made by Mr. Archie LaPointe, the Executive Director of one such program in the U.S., The National Assessment for Educational Progress (NAEP). In his address, Mr. LaPointe pointed out the degree to which arithmetic skills are acquired differently now than twenty years ago among nine-year olds, twelve-year olds, and fifteen-year olds. The NAEP represents all fifty states and all sixteen thousand school districts. For reasons of logistics and cost the sample is small and therefore cannot be used for major research functions. It would be difficult, for instance, to isolate different styles of classroom pedagogy from the use of hand-held calculators or computers as explanatory variables of achievement rises or declines. Such questions require special research projects on local, rather than national, populations. The virtue of a national assessment given on a regular basis is that it provides an unambiguous benchmark on the progress which a nation makes in teaching basic cognitive skills. A national assessment is particularly important for countries whose student population is on the rise, where enrollments are on the increase. In much of Africa, Latin America, and Asia, for example, it is common to assume that students are learning as much as they once did when the system was smaller. The assumption is often false. But how false? In which subjects? In which types of schools? In which parts of the country? These types of monitoring questions are essential for the managers of any educational system who wish to be informed.

(2) Measuring achievement across countries. Most OECD countries recognize that their economies are interdependent and, in many ways, competitive. Comparative statistics on trade, manufacturing, welfare, crime, public spending, and private investment, for example, are considered a necessity. These statistics are studied carefully by managers both in and outside of government.

This is no less true of education. In education, countries share statistics on enrollments, progression rates, expenditures, and, increasingly, academic achievement. Elementary- and secondary-school achievement is measured through commonly designed standardized tests of science, mathematics, and reading comprehension. One recommendation is that developing countries measure their own academic achievement against that of other countries.

The source for this comparative information is the International Association for the Evaluation of Educational Achievement (IEA), a non-government organization made up of professional institutes of education. Considerable effort has gone into the assurance that the standardized tests represent the formal and official curriculum of each country. No tests are administered unless each country's representative gives approval. The results, however, are public.

China has decided to participate in the IEA science study. The Chinese officials in charge made a presentation at the seminar in addition to the overall discussion led by Mr. John Keeves on results of previous IEA studies.

The dilemma of using standardized testing across countries is very simple. On the one hand, education is a local enterprise, with the purposes of schooling directed to preserving local culture, history, civic pride, and language. At the same time, however, educational systems are attempting to teach many common cognitive skills. Increasingly it is recognized that the ability to manipulate figures and interpret the written word are common goals of curricula at the same grade level in widely divergent countries. This commonality, moreover, pertains to wealthy and impoverished countries alike, to countries such as Sweden, which can afford to spend US$300 on each pupil each year for classroom reading materials and supplies, as well as to countries such as Bolivia, which can afford to spend only 1 percent of that per child.

So the question is why should a developing country engage in research on academic achievement along with a country which can spend one hundred times more per student. Aren't drastically lower levels of achievement inevitable?

The answer is yes and no. "Yes" with respect to mean achievement, but "no" with respect to school effectiveness. International studies of academic achievement have taught educational managers an important lesson. Mean achievement comparisons across countries for science or mathematics are essentially meaningless unless they are weighted by two factors. One factor is the level of monetary resources brought to bear on the classroom situation. The second factor is the percentage of the age cohort enrolled. Thus, although average mathematics scores at the twelfth-grade level in Germany are higher than in the U.S., the U.S. is educating a higher proportion of its eighteen-year-old citizens at the twelfth-grade level. Similarly although fourteen-year old students in Thailand have lower reading scores than fourteen-year olds in Japan, Japan is able to invest ten times the level of monetary resources per child in reading materials. School achievement is not judged by "Olympic records", but rather by performance standards in relation to the proportion enrolled and the level of resources brought to bear.

Results have been revealing. Some very effective school systems -- judged by these standards of resource handicaps -- have been found in developing countries, particularly in Asia. Regardless of the specific results at one time or in one subject, however, the main point is the function which such cross-national testing serves. It serves as a unique managerial device. It can point out the strengths and weaknesses of an educational system in relation to other systems. It can inform the managers of educational systems whether, by comparison to others, they are teaching certain skills too early or too late;[6] whether certain classroom equipment or pedagogical philosophies are effective; whether certain managerial innovations -- cross-age teaching, centralized learning centers, modular instruction -- are functional. In the field of education, countries are watching the process of innovation in ways very similar to those of other industries. International research on academic achievement is very much like measuring achievement over time in one important respect. It makes each country's search for excellence tangible.

Using Examinations to Improve Classroom Pedagogy. It is common thought that teachers "teach to the test," and to a large extent that is true. Managers have three choices with regard to this backwash effect of examinations. They can fight it; they can ignore it; or they can use it. The recommendation to developing countries is that they use it.

For many years a popular technique among educators has been to minimize these backwash effects by emphasizing other criteria of selection. This works. When selection is made on the basis of grades or on the basis of recommendations, teachers and students will pay less attention to examinations because the function of the examinations will be less important. By fighting the backwash effect of examinations, educational managers make the assumption that the effect, or its extent, is harmful to good teaching.

The second choice, that of ignoring backwash effects, is also common. Even when an examination board is part of a Ministry of Education, the management of the examination board is usually in the hands of examination specialists and not of educators. Examination specialists see their role as external to the education system. They see their function as providing a professionally designed product. Whether in or outside a Ministry of Education, little motivates a testing agency to analyze test results for any reason other than to improve the test.

Fighting or ignoring the backwash effect are not the only options. It is equally possible to use the examination system for purposeful pedagogical ends. The managers of education systems in some instances are able to allocate resources and staff time to provide analytic information to schools. One example of this was provided at this seminar by Mr. Anthony Somerset. His presentation included a description of the quite dramatic impact of a national newsletter. The newsletter contained examples of common errors and the incorrect cognitive reasoning which caused the errors in the first place. The newsletter had an immediate effect on classroom teaching. Quite soon afterward scores improved on that portion of the examination which tested those particular cognitive skills. Because of the educational care given to the analysis in the newsletter, pedagogical improvement resulted. The improvement came first in the better-managed school districts but later to the others as well. In sum, a nation's teaching improved in very tangible ways and very quickly. The function of selection provided the incentive. The educational analysis in the newsletter provided the means.

---

[6] OECD countries discovered that Chinese children were expected to acquire certain arithmetic functions in grade 2 which, in accordance with certain Western theories of child development, were not taught until grade 4.

Mr. Somerset's examples created a great deal of discussion and, in fact, confusion. Each country, and within China, each province, represented at the seminar claimed to have an examination feedback system. Confusion occurred because feedback systems can be organized at three different levels, with three levels of pedagogical sophistication and geographical specification.

The first level appeared to be very common. Results of examinations are reported in statistical terms which are often broken down to a region or province. Means and standard deviations for each test and, occasionally, for each test item are reported.

The second was the level described by Mr. Somerset. Means, standard deviations, and item analysis become the raw material for a pedagogical analysis of why errors occur in patterns. These patterns are then explained and, most important teaching methods are suggested on how to overcome errors. Thus the tone of the feedback mechanism in this second level is shifted from statistical terms to educational terms, the latter being the common language among teachers. The tone is shifted from one of negative sanctions as a result of public embarrassment caused by poor performance to one of detailed encouragement and specific suggestions for improvement. This second level is time consuming, expensive, and has only been obtained on large geographical units -- regions or provinces. This level of feedback mechanism is rare anywhere.

The third level is the educational analysis of selection-examination results school-by-school. Could such a luxury be afforded? In fact wealthy school districts in the U.S. maintain analytic profiles of each child, and monitor each child's progress on each objective within each subject. But these micro-analyses of performance patterns are used with assessment tests, not selection tests. Thus, the incentive for students and teachers to make improvements is significantly less. Nonetheless, the technique of computerizing day-to-day performance is affordable to some and is in use. This third level of feedback on selection-test results, however, remains thus far unattained.

Test Content and Format

Multiple Choice. In Nigeria, Indonesia, India, Brazil, China, and in other large and populous countries, the exigencies of economy and logistics will determine test format. For instance, China might begin to introduce multiple choice into its testing program and, as experience is gained and as the computers and optical-scanning equipment are acquired at both national and provincial levels, China might utilize multiple-choice formats only after the requisite technical experience has been gained in design. It is very easy to write a bad multiple-choice test item; it is not at all easy to design one well.

Since multiple-choice test formats are not necessarily a sign of good testing, a multiple-choice test is not therefore a sign of "modernity". Developing countries need not relinquish what common sense dictates: the written essay must remain. Just as the technical issues of designing multiple-choice formats are not trivial, so it must be mentioned that the requirements for standardizing essay responses are substantial. They are particularly complicated when large-scale testing is being conducted in languages which have not long been utilized for such purposes -- Kiswahili, Quecha, Indonesian, Filipino, Nepali, Creole. All languages have local dialects and terminological variations; all languages have stylistic alternatives. But in those languages which have long been used in large-scale testing -- Japanese, French, and English are examples -- traditions of standardization have been tested through time; they are widely acknowledged and widely understood in the school system. Attempting to standardize excellence in an essay when the tradition is only now being developed is an endeavor involving a certain amount of risk. At least with a multiple-choice test item the strengths and weaknesses are more visible. They are there for all to see. With an essay test item the difficulties are more hidden. The difficulties begin to emerge only after the item is set and the test taken. Developing countries using a national language which has not been extensively used for educational testing may be wise to shift to a multiple-choice format as rapidly as possible.

Scholastic Aptitude Testing in Developing Countries. Only one OECD country reports the successful and regular use of the SAT, and that is the United States. Its use in the United States, however, is to some extent determined by the style of educational governance rather than the virtues of the SAT itself. As many different curricula exist as school districts in the U.S., with each feeling very strongly about its own independence. This makes a national curriculum-specific selection examination difficult to market.

The SAT, by all admission, is difficult to design well. It requires a strong research base and a large supply of the technical skills which make up a research base. Reasons of cost require that past tests be kept private. Because of its privacy and non-applicability to specific curriculum, its feedback influence on the classroom is, by comparison, small. For each of these reasons the SAT is not recommended for developing countries.

The SAT has two virtues which should not be ignored by developing countries. In conjunction with curriculum-specific grades (or achievement tests) it is a reliable predictor of future academic performance. It is able to locate academic talent in impoverished schools where the opportunity to learn has been minimal. Large countries, such as China, might experiment with the SAT on a regular basis. In a decade or two, the research base required to support the SAT may be available and would not have to be imported. Moreover, the statistical data generated from SAT experiments often make good cross-checks on the validity parameters of achievement tests currently in use. Regular statistics on the predictive validity of the local SAT versus the national achievement test can be very useful information to maintain. Moreover, many design issues in which local experts should gain experience are pertinent to the SAT. Nevertheless, it is unlikely that the SAT would be of use in developing countries as a standardized mechanism of selection for higher education.

Skills to be Tested. A substantial degree of consensus existed at this seminar over a hierarchy of curriculum objectives, beginning with the skills of knowledge recall and ending with the skills of synthesis and/or evaluation. A display chart of this hierarchy can be found in Table 11.1. Even more remarkable was the common reference to a book on the subject by Mr. Benjamin Bloom.[7] This book was referred to by curriculum specialists and by test designers from China as well as from all five OECD countries.

These was little agreement beyond listing the hierarchical objectives and the types of examination questions that each level of objective implied. It was acknowledged that multiple-choice tests were easiest to construct at the lower end of the hierarchy. The most important issue was item distribution

---

[7] Bloom, Benjamin. Taxonomy of Educational Objectives: The Classification of Educational Goals: Handbook I. The Cognitive Domain. London: Longmans, 1956.

Table 11.1. Summary Chart Of Question Types

Question Type	Student Activity	Examples
Knowledge (remembering)	Recalling facts or observations. Recalling definitions.	1. Who? 2. What? 3. Where? 4. When? 5. Why? 6. Define (The word "ostinato"). 7. List
Comprehension (understanding)	Giving descriptions. Stating main ideas. Comparing.	1. Describe (what happened when we went to the concert). 2. What is the main idea (in this scene)? 3. How are (these two paintings) alike?
Application (solving)	Applying techniques and rules to solve problems that have a single correct answer.	1. If (Bill mixes yellow and blue paint) what (color will he get)? 2. Classify (these poems as ballads, sonnets, or odes).
Analysis (analyzing)	Identifying motives or causes. Making inferences. Finding evidence to support generalizations.	1. Why (did Aesop write fables)? 2. Now that we've studied this, what can we conclude about (pop art in America)? 3. What does this tell us about (the playwright's attitude towards war)? 4. What evidence can you find to support (the principle that air expands when heated)?
Synthesis (creating)	Solving problems. Making predictions. Producing original communications.	1. Can you think up (a title for this story)? 2. How can we solve (this dilemma)? 3. How can we improve (our pantomime)? 4. What will happen (if we add music here)?
Evaluation (judging)	Giving opinions about issues. Judging the validity of ideas. Judging the merit of problem-solutions. Judging the quality of art and other product.	1. Do you agree (with Kathy)? 2. Do you believe (that this is the best way to proceed)? 3. What is your opinion (on this matter)? 4. Would it be better (to do it this way)? 5. Which (painting) do you like?

across the hierarchy. What proportion of a selection examination should be dedicated to testing skills of knowledge recall? What proportion should test application skills? What proportion should test the skills of synthesis?

Mr. Somerset, who had the most experience in designing feedback mechanisms, dwelled at some length on the feedback requirements for moving a national examination from one with a high proportion of recall questions to one with a low proportion of recall questions. The order of magnitude for an example in which this actually occurred is displayed in Table 11.2. His main concern was that when such a shift takes place it must be made explicit to teachers. The success of the reform (as represented by the shift in the examination's balance) will depend upon the degree to which teachers help drive it.

Table 11.2. Kenya Primary Leaving Examination Test Items by Type and Year

	1973	(%)	1976
Descriptive	74		23
Explanatory	18		28
Observation	8		21
Reasoning	0		28

Professor Gui Shichun, the designer, made a presentation to the seminar on the newly-developed English proficiency test (EPT), now widely used in China. As with all test designers he was faced with the difficult decision on how to apportion the weight of the examination across the skill hierarchy. He and his colleagues chose to apportion it in what resembles a normal curve. This curve is displayed in Table 11.3.

Fifteen percent of the test was allocated to testing knowledge-recall skills.[8] Twenty-six percent was allocated to comprehension skills. The bulk of the test weight (72.5 percent) was allocated to the three medium-level skills, with proportionally smaller attention devoted to the extremes, both low and high.

Is this appropriate? Is this appropriate for developing countries in general or just for China? Is it appropriate for science and mathematics as well as for English as a foreign language? This seminar provided no single answer, although it was agreed that recall knowledge had to be tested and was comparatively easy to test. Nevertheless it was also agreed that it was not desirable to have a "nation of newts" selected on the basis of an ability to recall rather than because of creativity. In the absence of any formal standard, apportioning the weight of a test across the hierarchy in a normal-like curve appeared reasonable and may be reasonable to suggest for other tests and for other developing countries.

---

[8] This is an apportionment of test weight across skills, not test items. Different skills require different numbers of items and will take different lengths of time.

Table 11.3. China's English Proficiency Test (EPT) Items by Type, 1985

	(%)
Knowledge	15
Comprehension	26
Application	25
Analysis	21.5
Synthesis	12.5

Summary

In sum, the recommendations for developing countries which emerged from this meeting were relatively straightforward. An efficient educational system is an essential ingredient for maximizing national economic performance. Examinations are essential for the fair and efficient management of an education system. Information on achievement over time and in conjuction with other countries is also essential. In addition, good management requires a feedback system to schools, preferably at the level of specific suggestions for improvement.

These divergent functions of examinations and standardized testing cannot be developed in isolation from one another. They require coordination. In particular, they require coordination in the development of research capacity, equipment acquisition, and training programs. There are no single answers for developing countries on examination content or format, but logistical economies weigh heavily in favor of multiple choice. Nevertheless, each country should be aware of the variety of policies being developed in neighboring and in OECD countries on the issues of what kind of multiple choice, in what amount, and on what subject. Ultimately each developing country will have to maintain its own counsel.

V

SELECTED BIBLIOGRAPHY

Brimer, Alan Madaus, Chapman, George F.; Kellaghan, Thomas; and Wood, Robert. *Sources of Differences in School Achievement.* London: National Foundation of Education Research Publishing Company, 1978.

Brown, C. W., and Ghiselli, E. D. "Per Cent Increase in Proficiency Resulting from the Use of Selective Devices." *Journal of Applied Psychology* 37 (1953): 341-44.

Casella, Alexander. "Recent Developments in China's University Recruitment System." *China Quarterly* 62 (June 1975): 300.

Cronback, Lee J., and Gleser, G. C. *Selection Theory and Personnel Decisions* (2nd edition). Urbana: University of Illinois Press, 1965.

Elley, Warwick B., and Livingstone, Ian D. *External Examinations and Internal Assessments: Alternative Plans for Reform*. Wellington: New Zealand Council for Educational Research, 1972.

Furneaux, W. D. *The Chosen Few: An Examination of Some Aspects of University Selection in Britian*. London: Oxford University Press, 1961.

Furth, Dorotea. Selection and Equity "An International Viewpoint." *Comparative Education Review* 22 (June 1978): 260.

Heyneman, Stephen P., and Currie, Janice K. *Schooling, Academic Achievement and Occupational Attainment in a Non-Industrialized Society*. Washington D.C.: University Press of America, 1979.

Hoffmann, Banesh. *The Tyranny of Testing*. New York: Collier Books, 1962.

Hunter, John E., and Schmidt, Frank L. "Fitting People to Jobs: Implications of Personnel Selection for National Productivity." *Human Performance and Productivity*, ed. E. A. Fleishman. Hillsdale, N.J.: Erlbaum, 1982.

Klitgaard, Robert. *Choosing Elites*. New York: Basic Books, 1985.

Klitgaard, Robert. "Education on the Auction Bloc: An Admissions Fable." *Change* 15, no. 2 (1983): 44-47.

Klitgaard, Robert. *Elitism and Meritocracy in Developing Countries: Selection Policies for Higher Education*. Baltimore and London: Johns Hopkins University Press, 1986.

Klitgaard, Robert. "Identifying Exceptional Performers." *Policy Analysis* 4, no. 4 (1978): 529-47.

Lofstedt, Jan-Ingvar. *Chinese Educational Policy: Changes and Contradictions, 1949-79*. Atlantic Highlands, N.J: Humanities Press, 1980.

Lord, Frederic M., and Novick, Melvin R. *Statistical Theories of Mental Test Scores*. Reading, Mass.: Addison-Wesley, 1968.

Miyazaki, Ichisada, *China's Examination Hell: The Civil Service Examinations of Imperial China*. (Translated from Japanese by Conrad Schirokauer) New Haven: Yale University Press, 1976.

Ottobre, Frances, M. (editor). *Criteria for Awarding School Leaving Certificates: An International Discussion*. Oxford: Pergamon Press, 1979.

Pinera, Sebastian, and Selowsky, Marcelo. "The Optimal Ability-Education Mix and the Misallocation of Resources within Education Magnitude for Developing Countries." *Journal of Development Economics* 8 (1981): 111-31.

Purves, Alan C., and Levine, Daniel U. *Educational Policy and International Assessment: Implications of the IEA Surveys of Achievement*. Berkeley: McCutchan Publishing Company, 1975.

Riesman, David. "Educational Reform at Harvard: Meritocracy and Its Adversaries." in Lipset, Seymour Martin, and Riesman, David. *Education and Politics at Harvard*. New York: McGraw-Hill, 1975: 392.

Selowsky, Marcelo. "Women's Access to Schooling and the Value Added of the Educational System: An Application to Higher Education." *Women and Poverty in the Third World*. Marya Buminic, Margaret A. Lycette, and William Paul McGreevey, editors. Baltimore: Johns Hopkins University Press, 1983.

Spaulding, Jr. *Imperial Japan's Higher Civil Service Examinations*. Princeton: Princeton University Press, 1967.

Spaulding, Seth, and Kargorian, Arka. "Democratization of Higher Education through New Admissions Strategies: A Comparative Study of Theory and Practice." *UNESCO* ED-81/WS/4, February 1981.

Taylor, H. C., and Russell, J. T. "The Relationship of Validity Coefficients to the Practical Effectiveness of Tests in Selection: Discussion and Tables." *Journal of Applied Psychology* 23 (1939): 565-78.

Taylor, Robert. *China's Intellectual Dilemma: Politics and University Enrollment*. Vancouver: University of British Columbia Press, 1981.

Wechsler, Harold S. *The Qualified Student: A History of Selective College Admission In America*. New York: John Wiley & Sons, 1977.

Wigdoe, Alexandra K., and Garner, Wendell R., ed. *Ability Testing: Uses, Consequences, and Controversies*. Report of the Committee on Ability Testing, National Research Council, vol. 2. Washington, D.C.: National Academy Press, 1982.

Young, Michael. *The Rise of Meritocracy, 1940 - 2033*. Hammondsworth, England: Penguin, 1957.

DISTRIBUTORS OF WORLD BANK PUBLICATIONS

ARGENTINA
Carlos Hirsch, SRL
Galeria Guemes
Florida 165, 4th Floor-Ofc. 453/465
1333 Buenos Aires

AUSTRALIA, PAPUA NEW GUINEA, FIJI, SOLOMON ISLANDS, VANUATU, AND WESTERN SAMOA
Info-Line
Overseas Document Delivery
Box 506, GPO
Sydney, NSW 2001

AUSTRIA
Gerold and Co.
A-1011 Wien
Graben 31

BAHRAIN
MEMBR Information Services
P.O. Box 2750
Manama Town 317

BANGLADESH
Micro Industries Development Assistance Society (MIDAS)
G.P.O. Box 800
Dhaka

BELGIUM
Publications des Nations Unies
Av. du Roi 202
1060 Brussels

BRAZIL
Publicacoes Tecnicas Internacionais Ltda.
Rua Peixoto Gomide, 209
01409 Sao Paulo, SP

CANADA
Le Diffuseur
C.P. 85, 1501 Ampere Street
Boucherville, Quebec
J4B 5E6

COLOMBIA
Enlace Ltda.
Carrera 6 No. 51-21
Bogota D.E.

Apartado Aereo 4430
Cali, Valle

COSTA RICA
Libreria Trejos
Calle 11-13
Av. Fernandez Guell
San Jose

COTE D'IVOIRE
Entre d'Edition et de Diffusion Africaines (CEDA)
04 B.P. 541
Abidjan 04 Plateau

CYPRUS
MEMRB Information Services
P.O. Box 2098
Nicosia

DENMARK
SamfundsLitteratur
Rosenoerns Alle 11
DK-1970 Frederiksberg C.

DOMINICAN REPUBLIC
Editora Taller, C. por A.
Restauracion
Apdo. postal 2190
Santo Domingo

EGYPT, ARAB REPUBLIC OF
Al Ahram
Al Galaa Street
Cairo

FINLAND
Akateeminen Kirjakauppa
P.O. Box 128
SF-00101
Helsinki 10

FRANCE
World Bank Publications
66, avenue d'Iéna
75116 Paris

GERMANY, FEDERAL REPUBLIC OF
UNO-Verlag
Poppelsdorfer Alle 55
D-5300 Bonn 1

GREECE
KEME
24, Ippodamou Street
Athens-11635

GUATEMALA
Librerias Piedra Santa
Centro Cultural Piedra Santa
11 calle 6-50 zona 1
Guatemala City

HONG KONG, MACAU
Asia 2000 Ltd.
6 Fl., 146 Prince Edward Road, W.
Kowloon
Hong Kong

HUNGARY
Kultura
P.O. Box 139
1389 Budapest 62

INDIA
Allied Publishers Private Ltd.
751 Mount Road
Madras—600 002

15 J.N. Heredia Marg
Ballard Estate
Bombay—400 038

13/14 Asaf Ali Road
New Delhi—110 002

17 Chittaranjan Avenue
Calcutta—700 072

Jayadeva Hostel Building
5th Main Road Gandhinagar
Bangalore—560 009

3-5-1129 Kachiguda Cross Road
Hyderabad—500 027

Prarthana Flats, 2nd Floor
Near Thakore Baug, Navrangpura
Ahmedabad—380 009

Patiala House
16-A Ashok Marg
Lucknow—226 001

INDONESIA
Pt. Indira Limited
Jl. Sam Ratulangi 37
Jakarta Pusat
P.O. Box 181

IRELAND
TDC Publishers
12 North Frederick Street
Dublin 1

ISRAEL
The Jerusalem Post
The Jerusalem Post Building
P.O. Box 81
Romema Jerusalem 91000

ITALY
Licosa Commissionaria Sansoni SPA
Via Lamarmora 45
Casella Postale 552
50121 Florence

JAPAN
Eastern Book Service
37-3, Hongo 3-Chome, Bunkyo-ku 113
Tokyo

JORDAN
Jordan Center for Marketing Research
P.O. Box 3143
Jabal
Amman

KENYA
Africa Book Service (E.A.) Ltd.
P.O. Box 45245
Nairobi

KOREA, REPUBLIC OF
Pan Korea Book Corporation
P.O. Box 101, Kwangwhamun
Seoul

KUWAIT
MEMRB
P.O. Box 5465

MALAYSIA
University of Malaya Cooperative Bookshop Limited
P.O. Box 1127, Jalan Pantai Baru
Kuala Lumpur

MEXICO
INFOTEC
Apartado Postal 22-860
Col. PE/A Pobre
14060 Tlalpan, Mexico D.F.

MOROCCO
Societe d'Etudes Marketing Marocaine
2 Rue Moliere, Bd. d'Anfa
Casablanca

THE NETHERLANDS
InOr Publikaties
Noorderwal 38
7241 BL Lochem

NEW ZEALAND
Hills Library and Information Service
Private Bag
New Market
Auckland

NIGERIA
University Press Limited
Three Crowns Building Jericho
Private Mail Bag 5095
Ibadan

NORWAY
Tanum-Karl Johan, A.S.
P.O. Box 1177 Sentrum
Oslo 1

OMAN
MEMRB Information Services
P.O. Box 1613, Seeb Airport
Muscat

PAKISTAN
Mirza Book Agency
65, Shahrah-e-Quaid-e-Azam
P.O. Box No. 729
Lahore 3

PERU
Editorial Desarrollo SA
Apartado 3824
Lima

THE PHILIPPINES
National Book Store
701 Royal Avenue
Metro Manila

POLAND
ORPAN
Palac Kultury i Nauki
00-901 WARSZAWA

PORTUGAL
Liveria Portugal
Rua Do Carmo 70-74
1200 Lisbon

SAUDI ARABIA, QATAR
Jarir Book Store
P.O. Box 3196
Riyadh 11471

SINGAPORE, TAIWAN, BURMA, BRUNEI
Information Publications
Private, Ltd.
02-06 1st Fl., Pei-Fu Industrial
Bldg., 24 New Industrial Road
Singapore

SOUTH AFRICA
For single titles:
Oxford University Press Southern Africa
P.O. Box 1141
Cape Town 8000

For subscription orders:
International Subscription Service
P.O. Box 41095
Craighall
Johannesburg 2024

SPAIN
Mundi-Prensa Libros, S.A.
Castello 37
28001 Madrid

SRI LANKA AND THE MALDIVES
Lake House Bookshop
P.O. Box 244
100, Sir Chittampalam A. Gardiner Mawatha
Colombo 2

SWEDEN
For single titles:
ABCE Fritzes Kungl. Hovbokhandel
Regeringsgatan 12, Box 16356
S-103 27 Stockholm

For subscription orders:
Wennergren-Williams AB
Box 30004
S-104 25 Stockholm

SWITZERLAND
Librairie Payot
6 Rue Grenus
Case postal 381
CH 1211 Geneva 11

TANZANIA
Oxford University Press
P.O. Box 5299
Dar es Salaam

THAILAND
Central Department Store
306 Silom Road
Bangkok

TRINIDAD & TOBAGO, ANTIGUA, BARBUDA, BARBADOS, DOMINICA, GRENADA, GUYANA, JAMAICA, MONTSERRAT, ST. KITTS AND NEVIS, ST. LUCIA, ST. VINCENT & GRENADINES
Systematics Studies Unit
55 Eastern Main Road
Curepe
Trinidad, West Indies

TURKEY
Haset Kitapevi A.S.
469, Istiklal Caddesi
Beyoglu-Istanbul

UGANDA
Uganda Bookshop
P.O. Box 7145
Kampala

UNITED ARAB EMIRATES
MEMBR Gulf Co.
P.O. Box 6097
Sharjah

UNITED KINGDOM
Microinfo Ltd.
P.O. Box 3
Alton, Hampshire GU 34 2PG
England

VENEZUELA
Libreria del Este
Apdo. 60.337
Caracas 1060-A

YUGOSLAVIA
Jugoslovenska Knjiga
YU-11000 Belgrade Trg Republike

ZIMBABWE
Textbook Sales Pvt. Ltd.
Box 3799
Harare

RECENT WORLD BANK TECHNICAL PAPERS (continued)

No. 51. Wastewater Irrigation: Health Effects and Technical Solutions

No. 52. Urban Transit Systems: Guidelines for Examining Options

No. 53. Monitoring and Evaluating Urban Development Programs: A Handbook for Program Managers and Researchers

No. 54. A Manager's Guide to "Monitoring and Evaluating Urban Development Programs"

No. 55. Techniques for Assessing Industrial Hazards: A Manual

No. 56. Action-Planning Workshops for Development Management: Guidelines

No. 57. The Co-composing of Domestic Solid and Human Wastes

No. 58. Credit Guarantee Schemes for Small and Medium Enterprises

No. 59. World Nitrogen Survey

No. 60. Community Piped Water Supply Systems in Developing Countries: A Planning Manual

No. 61. Desertification in the Sahelian and Sudanian Zones of West Africa

No. 62. The Management of Cultural Property in World Bank-Assisted Projects: Archaeological, Historical, Religious, and Natural Unique Sites

No. 63. Financial Information for Management of a Development Finance Institution: Some Guidelines

No. 64. The Efficient Use of Water in Irrigation: Principles and Practices for Improving Irrigation in Arid and Semiarid Regions

No. 65. Management Contracts: Main Features and Design Issues

No. 66F. Preparation of Land Development Projects in Urban Areas

No. 67. Household Energy Handbook: An Interim Guide and Reference Manual

No. 68. Bus Services: Reducing Costs, Raising Standards

No. 69. Corrosion Protection of Pipelines Conveying Water and Wastewater

No. 70. Desertification Control and Renewable Resource Management in the Sahelian and Sudanian Zones of West Africa

No. 71. Reservoir Sedimentation: Impact, Extent, and Mitigation

No. 72. The Reduction and Control of Unaccounted-for Water: Working Guidelines

No. 73. Water Pollution Control: Guidelines for Project Planning and Financing

No. 74. Evaluating Traffic Capacity and Improvements to Road Geometry

No. 75. Small-Scale Mining: A Review of the Issues

No. 76. Industrial Minerals: A Technical Review

No. 77. Wastewater Management for Coastal Cities: Ocean Disposal Technologies

The World Bank

Headquarters
1818 H Street, N.W.
Washington, D.C. 20433, U.S.A.

Telephone: (202) 477-1234
Telex: WUI 64145 WORLDBANK
RCA 248423 WORLDBK
Cable Address: INTBAFRAD
WASHINGTONDC

European Office
66, avenue d'Iéna
75116 Paris, France

Telephone: (1) 47.23.54.21
Telex: 842-620628

Tokyo Office
Kokusai Building
1-1 Marunouchi 3-chome
Chiyoda-ku, Tokyo 100, Japan

Telephone: (03) 214-5001
Telex: 781-26838

Cover design by Bill Fraser

ISSN 0253-7494
ISBN 0-8213-0990-0